Fast-Rate PAC-Bayesian Generalization Bounds for Meta-Learning

Jiechao Guan [1 2] **Zhiwu Lu** [3 2]

Abstract

PAC-Bayesian error bounds provide a theoretical guarantee on the generalization abilities of meta-learning from training tasks to unseen tasks. However, it is still unclear how tight PAC-Bayesian bounds we can achieve for meta-learning. In this work, we propose a general PAC-Bayesian framework to cope with single-task learning and meta-learning uniformly. With this framework, we generalize the two tightest PAC-Bayesian bounds (i.e., kl-bound and Catoni-bound) from single-task learning to standard meta-learning, resulting in fast convergence rates for PAC-Bayesian meta-learners. By minimizing the derived two bounds, we develop two meta-learning algorithms for classification problems with deep neural networks. For regression problems, by setting Gibbs optimal posterior for each training task, we obtain the closed-form formula of the minimizer of our Catoni-bound, leading to an efficient Gibbs meta-learning algorithm. Although minimizing our kl-bound can not yield a closed-form solution, we show that it can be extended for analyzing the more challenging meta-learning setting where samples from different training tasks exhibit interdependencies. Experiments empirically show that our proposed meta-learning algorithms achieve competitive results with respect to latest works.

1. Introduction

Inspired by human beings' ability of utilizing past experience to efficiently learn a novel task, meta-learning, also referred to as *learning to learn* (Thrun & Pratt, 1998), has received much attention from the machine learning community in the last decade. The goal of meta-learning is thus to transfer the knowledge extracted from training tasks to unseen tasks for fast adaptation. Successful applications of such learning paradigm have been witnessed in computer vision (Snell et al., 2017; Ye et al., 2020), natural language processing (Wu et al., 2020; Chen et al., 2021), reinforcement learning (Li et al., 2021) and other related fields.

Apart from practical applications, the theoretical analysis of meta-learning has also been developed in the last decade. The pioneering work was provided by (Baxter, 2000), which first assumed that all learning tasks are independently and identically distributed (i.i.d.) from a task environment. Under this assumption, the following works investigated meta-learning from the standpoint of VC-theory (Maurer, 2009; Maurer et al., 2016) or algorithmic stability (Maurer, 2005; Chen et al., 2020). In recent years, there has also emerged an interest in studying meta-learning via PAC-Bayesian analysis (Pentina & Lampert, 2014; 2015; Amit & Meir, 2018). However, it is still unclear how tight PAC-Bayesian generalization bounds we can achieve for meta-learning.

This paper intends to tackle this issue and provides fast-rate PAC-Bayesian bounds for meta-learning. Our motivation is to generalize the two tightest PAC-Bayesian bounds, kl-bound and Catoni-bound, from the i.i.d. single-task learning setting (Maurer, 2004; Catoni, 2007), to the standard meta-learning setting where observations from different tasks are independent but not identically distributed. The main tool to achieve this goal is a useful lemma in (Berend & Tassa, 2010) that bounds the sum of independent bounded random variables with the sum of i.i.d. Bernoulli random variables. We can then apply the demonstration techniques in single-task learning to meta-learning, obtaining two fast-rate PAC-Bayesian bounds, which are still called kl-bound and Catoni-bound for convenience (Theorems 3-4). As a result, we unify the demonstration framework of PAC-Bayesian theory for single-task learning and meta-learning, and provide two up-to-date tightest bounds for meta-learners (see comparisons between different bounds in Table 1). By setting the derived two bounds as minimization objective, we also develop two bound-minimizing meta-learning algorithms for *classification* problems with deep neural networks.

Next, we give more theoretical results and applications of our derived two bounds. For the Catoni-bound, we show

[1]School of Information, Renmin University of China, Beijing, China [2]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China [3]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China. Correspondence to: Zhiwu Lu <luzhiwu@ruc.edu.cn>, Jiechao Guan <2014200990@ruc.edu.cn>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

how to obtain a closed-form solution to finding its minimum. Concretely, by choosing Gibbs posterior for each training task in this bound, we can yield an explicit form of the Gibbs optimal hyper-posterior, which is the training objective in PAC-Bayesian meta-learning setting. Directly approximating such closed-form hyper-posterior leads to an efficient meta-learning algorithm for *regression* problems. For the kl-bound, we extend it to the more challenging meta-learning setting where dependence exists among samples from different tasks. The approach undertaken to establish our result is based on the decomposition of the dependency graph of the training data (Ralaivola et al., 2010). The extended PAC-Bayesian kl-bound admits an analogous form of our previous one for the standard setting, up to a multiplicative factor that represents the degree of interdependence within the dependency graph. This bound thus reveals the impact of data dependence for meta-learning. Finally, we conduct experiments on several benchmarks. The empirical results show that our proposed meta-learning algorithms achieve competitive performance with respect to latest works.

Our main contributions are summarized as follows:

(1) We propose a unified framework that can generalize PAC-Bayesian analysis from single-task learning to meta-learning. The fast-rate PAC-Bayesian kl-bound and Catoni-bound are hence derived for meta-learning, followed by two bound-minimizing classification algorithms.

(2) By setting Gibbs optimal posterior for each training task in our Catoni-bound, we obtain the explicit form of the optimal hyper-posterior. We thus develop an efficient regression algorithm that only needs to approximate the Gibbs optimal hyper-posterior, instead of learning posterior for each task and hyper-posterior simultaneously like previous methods.

(3) We extend our kl-bound to the meta-learning setting where tasks are dependent. To the best of our knowledge, this is the first PAC-Bayesian bound for meta-learning with data from different tasks exhibiting interdependencies.

(4) Experiment results on both classification and regression problems empirically validate the effectiveness of our theoretical analysis for meta-learning, especially the tightness and applicability of our PAC-Bayesian Catoni-bound.

2. Related Work

PAC-Bayesian Theory. The first PAC-Bayesian generalization bound was provided by (McAllester, 1998). (Seeger, 2002) developed such theory and obtained the tightest PAC-Bayesian kl-bound. (Germain et al., 2009) provided a general method to demonstrate a PAC-Bayesian bound. We need to point out that in the above works, the i.i.d. assumption is necessary to obtain the tightest kl-bound. Besides that, (Catoni, 2007) yielded a generalization bound of fast rate by just assuming the independence of observations. Until now, the kl-bound in (Maurer, 2004) and Catoni's bound are still among the tightest PAC-Bayesian bounds. More explanations for the similarities between these tight bounds can be found in (Audibert, 2010) [Chapter 2]. Other works extended PAC-Bayesian bounds to general cases, such as martingale setting (Seldin et al., 2012), heavy-tailed data (Alquier & Guedj, 2018), non-i.i.d. data (Ralaivola et al., 2010) or unbounded loss functions (Holland, 2019; Haddouche et al., 2021). In this work, we focus on bounded loss function. We first propose the generalized PAC-Bayesian kl-bound and Catoni-bound for independent (not necessarily identically distributed) random variables (Theorem 2). Then, we show how to apply them to derive fast-rate PAC-Bayesian generalization error bounds for meta-learning.

Meta-Learning Theory. The first systematic analysis of meta-learning theory was proposed by (Baxter, 2000), which assumed that the data distributions of different tasks are i.i.d. sampled from the same task environment. Baxter then gave a covering number based generalization bound for meta-learning. After that, (Maurer, 2005) investigated meta-learning theory from the perspective of algorithm stability and proposed a new indicator, called transfer risk, to measure the performance of a meta-learning algorithm. Other works followed this line and intended to provide a tighter bound on the transfer risk (Maurer, 2009; Denevi et al., 2018; Chen et al., 2020). In this work, we study meta-learning from the perspective of PAC-Bayesian analysis.

PAC-Bayesian Meta-Learning Theory. The pioneering work of PAC-Bayesian meta-learning theory was proposed by (Pentina & Lampert, 2014). They adopted the task environment notation from Baxter and proposed the concept of *hyper-posterior*. The hyper-posterior is trained over the observed tasks and generates an informative prior when encountering a novel task for fast adaptation. The PAC-Bayesian bound on the transfer risk (of the learned hyper-posterior) is always composed of three parts: empirical multi-task error, environment-level complexity and task-level complexity (see Table 1). (Pentina & Lampert, 2015) further generalized PAC-Bayesian meta-learning theory to the non-i.i.d. cases where different tasks are dependent or the task environment is changing. (Amit & Meir, 2018) provided a new PAC-Bayesian bound and applied their theoretical results to the optimization of deep neural networks. (Rothfuss et al., 2021) generalized the PAC-Bayesian meta-learning theorem to the unbounded loss and derived the PAC-optimal hyper-posterior. (Farid & Majumdar, 2021) studied meta-learning with both PAC-Bayes and uniform stability analysis. However, both the environment-level complexity and the task-level complexity in these PAC-Bayesian bounds suffered from slow convergence rates (e.g., the environment-level and task-level complexities in (Pentina & Lampert, 2015) are of order $O(\frac{1}{\sqrt{n}})$ and $O(\frac{1}{n\sqrt{m}})$, respectively, with n training tasks and m samples per task). In this work, we

Table 1. Different PAC-Bayesian meta-learning bounds on $er(\mathcal{Q})$. **Bound = Empirical Error + Environment-level Complexity + Task-level Complexity**. n is the number of training tasks. m is the sample size per task. $\mathcal{P}, \mathcal{Q} \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{H}))$ are the hyper-prior and hyper-posterior respectively. $P, Q_i = Q(S_i, P) \in \mathcal{M}_1(\mathcal{H})$ are the prior and the posterior for the i-th training task. In our Catoni-bound, the constant $C > 1$. Explicit forms of different bounds are given in Table 2 of the Appendix.

Classical Bounds	Empirical Error	Environment-Level Complexity	Task-Level Complexity
(Pentina & Lampert, 2014)	$\widehat{er}(\mathcal{Q})$	$O\big(\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})}{\sqrt{n}}\big)$	$O\big(\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})+\sum_{i=1}^{n}\mathbf{E}_{P\sim\mathcal{Q}}\mathcal{K}(Q_i,P)}{n\sqrt{m}}+\frac{1}{\sqrt{m}}\big)$
(Amit & Meir, 2018)	$\widehat{er}(\mathcal{Q})$	$O\big(\sqrt{\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})+\ln n}{n}}\big)$	$O\big(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\mathcal{K}(Q_i,P)+\ln(2nm)}{m}}\big)$
(Rothfuss et al., 2021)	$\widehat{er}(\mathcal{Q})$	$O\big(\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})}{\sqrt{n}}\big)$	$O\big(\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})+\sum_{i=1}^{n}\mathbf{E}_{P\sim\mathcal{Q}}\mathcal{K}(Q_i,P)}{n\sqrt{m}}+\frac{1}{\sqrt{n}}\big)$
kl-bound (ours)	$\widehat{er}(\mathcal{Q})$	$O\big(\sqrt{\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})+\ln\sqrt{n}}{n}}\big)$	$O\big(\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\sum_{i=1}^{n}\mathcal{K}(Q_i,P)+\ln\sqrt{nm}}{mn}\big)$
Catoni-bound (ours)	$C\widehat{er}(\mathcal{Q})$	$O\big(\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})}{n}\big)$	$O\big(\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\sum_{i=1}^{n}\mathcal{K}(Q_i,P)}{mn}\big)$

derive two fast-rate PAC-Bayesian bounds for meta-learning (Theorem 3-4). Specifically, the task-level complexities in our kl-bound and Catoni-bound have a rate of $O(\frac{1}{nm})$. The environment-level complexity in our Catoni-bound also have a fast rate of $O(\frac{1}{n})$ (see Table 1 for detailed comparisons). Furthermore, we generalize our kl-bound to the meta-learning setting of dependent observations. Note that although (Pentina & Lampert, 2015) claimed that they derived a bound for dependent tasks, they still assumed the independence of the samples from different tasks. In contrast, we provide a more general PAC-Bayesian bound for meta-learning with dependent samples from different tasks, and show that this bound is much tighter than that in (Pentina & Lampert, 2015) (see discussion below Theorem 6).

3. Preliminary

A supervised learning problem is characterized by the sample space $\mathcal{Z}$, a hypothesis space $\mathcal{H}$, and a loss function $l : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$, where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ is the product space of input space $\mathcal{X}$ and label space $\mathcal{Y}$. We assume that l is $[0, 1]$-valued. Actually, if the loss function l is bounded and locates in the interval $[0, M]$ ($M > 0$), we can use the rescaling technique and focus on the $[0, 1]$-valued loss l/M. Let $[K]$ denote the set $\{1, ..., K\}$, for any integer K. $\mathcal{M}_1(A)$ denotes the set of probability measures over the set A. Throughout the paper, we ignore measurability issues.

3.1. PAC-Bayesian Theory for Single-Task Learning

In PAC-Bayesian single-task learning setting, a task is characterized by an unknown distribution D over the space $\mathcal{Z}$, from which a size-m sample $S = \{z_i\}_{i=1}^m$ is provided, with each z_i drawn i.i.d. from D. For any hypothesis $h \in \mathcal{H}$, denote by $er(h, D) \triangleq \mathbf{E}_{z\sim D} l(h, z)$ its expected error over D and by $\widehat{er}(h, S) \triangleq \frac{1}{m}\sum_{i=1}^{m} l(h, z_i)$ its empirical error over S. In the PAC-Bayesian framework, the goal is to output a posterior $Q = Q(S, P) \in \mathcal{M}_1(\mathcal{H})$ by training an algorithm with the sample S and the prior $P \in \mathcal{M}_1(\mathcal{H})$ as input. The expected error and empirical error of a randomized classifier associated with distribution Q are defined as $er(Q, D) \triangleq \mathbf{E}_{h\sim Q} er(h, D)$ and $\widehat{er}(Q, S) \triangleq \mathbf{E}_{h\sim Q}\widehat{er}(h, S)$, respectively. Denote the KL-divergence between distributions Q and P by $\mathcal{K}(Q, P) = \mathbf{E}_{h\sim Q} \ln \frac{\mathrm{d}Q}{\mathrm{d}P}$, where $\frac{\mathrm{d}Q}{\mathrm{d}P}$ represents the Radon-Nikodym derivative of Q with respect to P. Denote the relative entropy between the Bernoulli random variables with success rate p and q by $\mathrm{kl}(p, q) = p \ln \frac{p}{q} + (1 - p) \ln \frac{1-p}{1-q}$. Then the PAC-Bayesian kl-bound and Catoni-bound for single-task learning are:

Theorem 1 *(Germain et al., 2009) [Corollary 2.1-2.2] Let l be the binary-valued misclassification loss. For any fixed prior $P \in \mathcal{M}_1(\mathcal{H})$, any data distribution D over $\mathcal{Z}$, any $\delta > 0$, any positive constant $\lambda > 0$, with probability at least $1 - \delta$ over the draw of i.i.d. sample $S \sim D^m$, the following two inequalities hold for any posterior $Q \in \mathcal{M}_1(\mathcal{H})$:*

$$\mathrm{kl}(\widehat{er}(Q, S), er(Q, D)) \le \frac{\mathcal{K}(Q, P) + \ln \frac{2\sqrt{m}}{\delta}}{m},$$

$$er(Q, D) \le \frac{\lambda}{m(1-\mathrm{e}^{-\frac{\lambda}{m}})}\widehat{er}(Q, S) + \frac{\mathcal{K}(Q, P)+\ln(1/\delta)}{m(1-\mathrm{e}^{-\frac{\lambda}{m}})}.$$

3.2. PAC-Bayesian Theory for Meta-Learning

In PAC-Bayesian meta-learning setting, the learner is given n different training tasks, each of which is associated with a data distribution $D_i (1 \le i \le n)$ over the sample space $\mathcal{Z}$ (i.e. $D_i \in \mathcal{M}_1(\mathcal{Z})$). In each task, the size-m i.i.d. training sample $S_i = \{z_{ij}\}_{j=1}^m \sim D_i^m$ is provided. To develop meta-learning theory, we adopt the *task environment* concept proposed by (Baxter, 2000). Specifically, we assume that the n different data distributions $\{D_i\}_{i=1}^n$ are sampled from the same distribution, referred to as environment τ. Thus, the environment τ can be regarded as a probability measure over the set of all distributions (i.e. $\tau \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{Z}))$).

Next, we follow the PAC-Bayesian meta-learning framework proposed by (Pentina & Lampert, 2014). We regard the prior $P \in \mathcal{H}$ for each training task as a random variable. P

is sampled randomly from a distribution called *hyper-prior* $\mathcal{P} \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{H}))$ given before seeing the n training tasks. The goal of meta-learning is to output a *hyper-posterior* $\mathcal{Q} \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{H}))$ by training a meta-learning algorithm with hyper-prior $\mathcal{P}$ and datasets $\{S_i\}_{i=1}^n$ as input. When encountering the new task, a new prior will be sampled from the learned hyper-posterior $\mathcal{Q}$ that contains the information of the n training tasks for fast adaptation. Under the PAC-Bayesian framework, choosing an informative prior for a new task can achieve a tighter generalization bound and better performance (Catoni, 2007). Since different tasks are sampled from the same environment and share some similarities, the informative prior drawn from the learned hyper-posterior is expected to adapt to the new task quickly. Formally, the quality of the learned hyper-posterior $\mathcal{Q}$ can be measured by the *transfer risk* (Maurer, 2009; Pentina & Lampert, 2014) on the new data distribution D sampled independently from the same task environment τ:

$$er(\mathcal{Q}) \triangleq \mathbf{E}_{P\sim\mathcal{Q}}\mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} er(Q(S,P),D). \quad (1)$$

Notice that we can not minimize the transfer risk directly since the environment τ and the sampled probability measure D are unknown. Instead, we will choose to minimize the following *empirical multi-task risk* over the n training tasks $\widehat{er}(\mathcal{Q})$ to obtain a high-quality hyper-posterior $\mathcal{Q}$:

$$\widehat{er}(\mathcal{Q}) \triangleq \mathbf{E}_{P\sim\mathcal{Q}}\frac{1}{n}\sum_{i=1}^{n}\widehat{er}(Q(S_i,P),S_i). \quad (2)$$

Furthermore, to obtain a PAC-Bayesian bound for $er(\mathcal{Q})$, we also consider the following *expected multi-task risk*:

$$\widetilde{er}(\mathcal{Q}) \triangleq \mathbf{E}_{P\sim\mathcal{Q}}1/n\sum_{i=1}^{n} er(Q(S_i,P),D_i). \quad (3)$$

We will write $Q(S_i,P)$ as Q_i for abbreviation when the context is clear. The notations of PAC-Bayesian single-task learning and meta learning are listed in Table 6 in Appendix H for convenient reference. The PAC-Bayesian meta-learning theory thus will give an upper bound on the transfer risk $er(\mathcal{Q})$ of the hyper-posterior $\mathcal{Q}$ according to its empirical multi-task risk $\widehat{er}(\mathcal{Q})$ or its expected multi-task risk $\widetilde{er}(\mathcal{Q})$. Such results will be detailed in the next section.

4. Theoretical Results

We provide fast-rate PAC-Bayesian bounds for meta-learning in two common scenarios: (1) All samples and all tasks are independent (we also show how to derive the Gibbs optimal hyper-posterior by minimizing our Catoni-bound). (2) Samples from different tasks are dependent.

4.1. PAC-Bayesian Bounds for Meta-Learning with Independent Samples

To give fast-rate PAC-Bayesian bounds for meta-learning, we need to choose convex function $\mathcal{D}(p,q)$ and then bound the moment generating function (MGF) of $\mathcal{D}(p,q)$ (i.e., $\mathbf{E}\exp\{\mathcal{D}(er(\mathcal{Q}),\widetilde{er}(\mathcal{Q}))\}$ and $\mathbf{E}\exp\{\mathcal{D}(\widetilde{er}(\mathcal{Q}),\widehat{er}(\mathcal{Q}))\}$). Almost all existing works set $\mathcal{D}(p,q) = p-q$, apply Hoeffding's lemma to bound the MGF of $\mathcal{D}(p,q)$, and finally obtain a PAC-Bayesian meta learning bound of $O(1/t + t/K)(\forall t>0)$, which suffers a slow convergence rate of $O(1/\sqrt{K})$ $(K>0)$. In contrast, we set $\mathcal{D}(p,q)$ as $\mathrm{kl}(q,p)$ or [1] $\Phi_{\frac{\lambda}{K}}(p) - q$, $(\lambda>0)$, as what we do to obtain the PAC-Bayesian kl-bound and Catoni-bound in single-task learning. However, since $\widetilde{er}(\mathcal{Q})$ and $\widehat{er}(\mathcal{Q})$ are the summations of independent $[0,1]$-valued random variables (not i.i.d. $\{0,1\}$-valued ones as in Theorem 1), we can not directly apply the results in Theorem 1 to bound the MGF of $\mathcal{D}(p,q)$. To overcome this challenge, we use the following lemma to bound the expectation of the function of the sum of independent $[0,1]$-valued random variables (rvs) with the expectation of the function of the sum of i.i.d. $\{0,1\}$-valued ones. Such result is originated from (Berend & Tassa, 2010), and more explanations can be found in Appendix A.

Lemma 1 *Let $\{\xi_k\}_{k=1}^K$ be a sequence of independent random variables with $P(0\le\xi_k\le 1)=1$, and $\{\eta_k\}_{k=1}^K$ be a sequence of i.i.d. Bernoulli random variables with $\mathbf{E}\eta_k = K^{-1}(\sum_{k=1}^K \mathbf{E}\xi_k)$. Then for any convex function g,*

$$\mathbf{E}g(\frac{1}{K}\sum_{k=1}^{K}\xi_k) \le \mathbf{E}g(\frac{1}{K}\sum_{k=1}^{K}\eta_k).$$

With such lemma, it is more convenient for us to derive generalized PAC-Bayesian kl-bound or Catoni-bound for independent $[0,1]$-valued random variables as follow.

Theorem 2 *Let $\mathcal{F}$ be a set of random variables f. Let $\mathcal{S} = \{\xi_k\}_{k=1}^K$ be a sequence of random variables with each component ξ_k ($k\in[K]$) drawn independently according to the measure μ_k over the set A_k. Let $R(f) = \frac{1}{K}\sum_{k=1}^K \mathbf{E}_{\xi_k} g_k(f,\xi_k)$, $r(f) = \frac{1}{K}\sum_{k=1}^K g_k(f,\xi_k)$, where $g_k : \mathcal{F}\times A_k \to [0,1]$ is a bounded function. Denote $\mathbf{E}_{f\sim\rho}(R(f)), \mathbf{E}_{f\sim\rho}(r(f))$ by $\rho(R), \rho(r)$ respectively. Then $\forall\delta>0, \lambda>0$, for any predefined distribution $\pi\in\mathcal{M}_1(\mathcal{F})$, with probability at least $1-\delta$ over the draw of $\mathcal{S}$, the following two inequalities hold for any measure ρ over $\mathcal{F}$:*

$$\mathrm{kl}(\rho(r),\rho(R)) \le \frac{\mathcal{K}(\rho,\pi)+\ln(2\sqrt{K}/\delta)}{K},$$

$$\rho(R) \le \frac{\lambda\rho(r)}{K(1-\mathrm{e}^{-\frac{\lambda}{K}})} + \frac{\mathcal{K}(\rho,\pi)+\ln(1/\delta)}{K(1-\mathrm{e}^{-\frac{\lambda}{K}})}.$$

Proof Sketch. Note that $\mathcal{D}(\rho(R),\rho(r)) \le \frac{1}{\lambda}\big[\mathcal{K}(\rho,\pi) + \ln\mathbf{E}_{f\sim\pi}\mathbf{E}_{\mathcal{S}}\mathrm{e}^{\lambda\mathcal{D}(R(f),r(f))}/\delta\big]$ holds with high probability for any convex function $\mathcal{D}(\cdot,\cdot)$. With Lemma 1 we can bound $\mathbf{E}_{\mathcal{S}}\mathrm{e}^{\lambda\mathcal{D}(R(f),r(f))}$ with the MGF of convex function of the sum of i.i.d. Bernoulli rvs. Setting $\mathcal{D}(p,q)$ as $\mathrm{kl}(q,p)$ or $\Phi_{\frac{\lambda}{K}}(p)-q$, and using Theorem 1 finish the proof. ■

[1] $\Phi_a(p) = -a^{-1}\ln\{1-[1-\exp(-a)]p\}$, $a\in\mathbb{R}$, $p\in[0,1]$

Setting $\mathcal{F}=\mathcal{H}, g_k=l, A_k=\mathcal{Z}$, we can recover the result in Theorem 1. As shown in (Maurer, 2004) [Equation (2)], the PAC-Bayesian kl-bound in the right hand side (RHS) of the above first inequality gives the optimal order in K. Applying Pinsker's inequality $\mathrm{kl}(p,q) \geq 2(p-q)^2$, we can obtain a bound on the deviation $|\rho(R)-\rho(r)|$ with a slow convergence rate of $O(1/\sqrt{K})$. Further, applying the stronger version $\mathrm{kl}(p,q) \geq (p-q)^2/(2q)$ when $q>p$ gives a fast convergence rate of $O(\log K/K)$ for generalization bound. Such analysis leads to our fast-rate PAC-Bayesian kl-bound for meta-learning as below.

Theorem 3 *For any predefined hyper-prior $\mathcal{P}$, with probability at least $1-\delta$ over the draw of the training sample $\{S_i\}_{i=1}^n$, the following holds for any hyper-posterior $\mathcal{Q}$:*

$$er(\mathcal{Q}) \leq \widehat{er}(\mathcal{Q}) + \sqrt{\frac{\mathcal{K}(\mathcal{Q},\mathcal{P}) + \ln\frac{2\sqrt{n}}{\delta}}{2n}} + \sqrt{\frac{2\Delta\widehat{er}(\mathcal{Q})}{mn}} + \frac{2\Delta}{mn},$$

where $\Delta = \mathcal{K}(\mathcal{Q},\mathcal{P}) + \mathbf{E}_{P\sim\mathcal{Q}}\sum_{i=1}^n \mathcal{K}(Q_i,P) + \ln\frac{2\sqrt{mn}}{\delta}$.

To prove the above result, applying the kl-bound in Theorem 2 to bound $\mathrm{kl}(er(\mathcal{Q}),\widetilde{er}(\mathcal{Q}))$ and $\mathrm{kl}(\widetilde{er}(\mathcal{Q}),\widehat{er}(\mathcal{Q}))$ respectively and then using the union bound can obtain an upper bound on $er(\mathcal{Q})$. As shown in (Pentina & Lampert, 2014; Amit & Meir, 2018), the proof of the bound for meta-learning is always divided into two parts: bounding the deviations $er(\mathcal{Q})-\widetilde{er}(\mathcal{Q})$ and $\widetilde{er}(\mathcal{Q})-\widehat{er}(\mathcal{Q})$ respectively. The contribution of our Theorem 3 lies in the task-level complexity term $\sqrt{\frac{2\Delta\widehat{er}(\mathcal{Q})}{mn}}+\frac{2\Delta}{mn}$ on the deviation $\widetilde{er}(\mathcal{Q})-\widehat{er}(\mathcal{Q})$. When the empirical error $\widehat{er}(\mathcal{Q})$ is close to zero, such bound has an magnitude of $O(\frac{\ln(mn)}{mn})$, which is tighter than the bound of $O(\frac{1}{n\sqrt{m}})$ in (Pentina & Lampert, 2014, Theorem 1) when $n << m$. Besides that, the environment-level complexity of our derived kl-bound $O(\sqrt{\frac{\ln\sqrt{n}}{n}})$ halves the logarithmic dependence of n in the numerator, while it is $O(\sqrt{\frac{\ln n}{n}})$ in the bound of (Amit & Meir, 2018). Moreover, using the Catoni-bound in Theorem 2, we can achieve our tighter PAC-Bayesian Catoni-bound for meta-learning.

Theorem 4 *For any predefined hyper-prior $\mathcal{P}$, any $\delta\in(0,1)$, any $C_1,C_2>1$, with probability at least $1-\delta$ over the draw of the training sample $\{S_i\}_{i=1}^n$, the following holds for any hyper-posterior $\mathcal{Q}$:*

$$er(\mathcal{Q}) \leq \frac{C_1C_2\ln C_1\ln C_2}{(C_1-1)(C_2-1)}\widehat{er}(\mathcal{Q}) + \frac{C_1\big(\mathcal{K}(\mathcal{Q},\mathcal{P})+\ln(2/\delta)\big)}{n(C_1-1)} + \frac{C_1C_2\ln C_1\big(\mathcal{K}(\mathcal{Q},\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\sum_{i=1}^n\mathcal{K}(Q_i,P)+\ln(2/\delta)\big)}{(C_1-1)(C_2-1)nm}.$$

The proof strategy is also to bound the deviations $er(\mathcal{Q})-\widetilde{er}(\mathcal{Q})$ and $\widetilde{er}(\mathcal{Q})-\widehat{er}(\mathcal{Q})$ respectively. If we suppress the KL-complexities in the numerator, the bound on $er(\mathcal{Q})-\widetilde{er}(\mathcal{Q})$ has an order of $O(\frac{1}{n})$ and the bound on $\widetilde{er}(\mathcal{Q})-\widehat{er}(\mathcal{Q})$ has an order of $O(\frac{1}{mn})$. Both bounds have a fast convergence rate w.r.t. the number of their observations. Therefore, Theorem 4 provides the tightest PAC-Bayesian bound for standard meta-learning in this paper. Setting the above kl-bound and Catoni-bound as objective functions can lead to two meta-learning algorithms for classification problems, which will be detailed in the experiment section.

4.2. Optimizing PAC-Bayesian Catoni-Bound with Gibbs Optimal Hyper-posterior

In this subsection, we show how to utilize our Catoni-bound to obtain the closed-form formula of the Gibbs optimal hyper-posterior. As shown in (Zhang, 2006) [Equation (5)], we can derive an explicit form of the optimal posterior ρ^* for $\min_\rho\{\beta\rho(r)+\mathcal{K}(\rho,\pi)\}$, where $\beta\in\mathbb{R}$. The obtained posterior is called Gibbs optimal posterior (Catoni, 2007). Next, we apply such strategy to the minimization of our PAC-Bayesian Catoni-bound for meta-learning in Theorem 4 to establish the explicit form of the Gibbs optimal hyper-posterior. We first give a corollary of Theorem 4 by choosing the Gibbs optimal posterior for each training task.

Corollary 1 *$\forall i\in[n]$, any prior $P\in\mathcal{M}_1(\mathcal{H})$, any training data $\{S_i\}_{i=1}^n$, let Q_i^* be the Gibbs optimal posterior such that $\frac{\mathrm{d}Q_i^*}{\mathrm{d}P} = \exp\{-m\widehat{er}(h,S_i)\}/Z(S_i,P)$, where $Z(S_i,P)=\int_{\mathcal{H}}\mathrm{e}^{-m\widehat{er}(h,S_i)}\mathrm{d}P(h)$ is a normalization constant. Then $\forall\delta>0, C_1>1$, with probability at least $1-\delta$ over the draw of training datasets $\{S_i\}_{i=1}^n$, the following holds for any hyper-posterior $\mathcal{Q}$:*

$$er(\mathcal{Q}) \leq \frac{\mathrm{e}C_1\ln C_1}{(C_1-1)(\mathrm{e}-1)}\mathbf{E}_{P\sim\mathcal{Q}}\frac{-1}{nm}\sum_{i=1}^n[\ln Z(S_i,P)] + \frac{C_1\big(\mathcal{K}(\mathcal{Q},\mathcal{P})+\ln\frac{2}{\delta}\big)}{n(C_1-1)} + \frac{\mathrm{e}C_1\ln C_1\big(\mathcal{K}(\mathcal{Q},\mathcal{P})+\ln\frac{2}{\delta}\big)}{nm(C_1-1)(\mathrm{e}-1)}.$$

An analogous result that is also derived by setting Gibbs optimal posterior for each task can be found in (Rothfuss et al., 2021) [Corollary 1]. Omitting the empirical error part, both Rothfuss's bound and our bound share the same order of $O\big((\frac{1}{n}+\frac{1}{mn})\mathcal{K}(\mathcal{Q},\mathcal{P})\big)$. However, Rothfuss's bound has an extra constant (i.e., $1/8$) that can not vanish with the increase of the size of training samples. Therefore, our generalization bound has a better asymptotic behaviour. Next we can obtain the explicit form of Gibbs optimal hyper-posterior by minimizing the RHS of the inequality in Corollary 1.

Corollary 2 *(Gibbs Optimal Hyper-posterior) For any hyper-prior $\mathcal{P}\in\mathcal{M}_1(\mathcal{M}_1(\mathcal{H}))$ and any training datasets $\{S_i\}_{i=1}^n$, the hyper-posterior $\mathcal{Q}\in\mathcal{M}_1(\mathcal{M}_1(\mathcal{H}))$ that minimizes the PAC-Bayesian meta-learning bound in Corollary 1 has the following explicit form:*

$$\frac{\mathrm{d}\mathcal{Q}^*}{\mathrm{d}\mathcal{P}}(P) = \exp\{\frac{\beta}{nm}\sum_{i=1}^n\ln Z(S_i,P)\}/Z(\mathbf{S},\mathcal{P}),$$

where $\beta = \frac{eC_1 \ln C_1}{(C_1-1)(e-1)\alpha}$, $\alpha = \frac{eC_1 \ln C_1}{nm(C_1-1)(e-1)} + \frac{C_1}{n(C_1-1)}$, $Z(\mathbf{S}, \mathcal{P}) = \int_{\mathcal{M}_1(\mathcal{H})} \exp\{\frac{\beta}{nm} \sum_{i=1}^{n} \ln Z(S_i, P)\} \mathrm{d}\mathcal{P}(P)$ *is a normalization constant.*

As pointed out by (Rothfuss et al., 2021) [Proposition 1], the explicit form of the Gibbs optimal hyper-posterior in Corollary 2 makes it much easier to optimize the meta-learning bound in Corollary 1, than to directly optimize the Catoni-bound in Theorem 4 which needs to learn $\mathcal{Q}$ and $\{Q_i\}_{i=1}^n$ simultaneously. We will use statistical inference methods to approximate the above $\mathcal{Q}^*$ and thus develop an efficient Gibbs optimal hyper-posterior (GOHP) algorithm for regression problems in the experiment section.

4.3. PAC-Bayesian kl-Bound for Meta-Learning with Dependent Samples

In this subsection, we consider the meta-learning setting where dependence exists between different tasks and between different samples. Our strategy is to split the dependent random variables into different groups of independent random variables. Such splitting strategy is originated from (Hoeffding, 1963) and typically developed by (Janson, 2004) with graph decomposition techniques. We introduce two significant concepts about *Dependence Graph* that help us analyze the meta-learning setting with dependent samples.

Definition 1 *(Dependence Graph) Let* $\mathcal{S} = \{\xi_1, \ldots, \xi_K\}$ *be a set of* K *random variables. The dependence graph* $\Gamma(\mathcal{S}) = (V, E)$ *of* $\mathcal{S}$ *is such that*

- *the set of vertices* V *of* $\Gamma(\mathcal{S})$ *is* $V = [K]\}$.
- $(i, j) \notin E$ *(i.e., there is no edge between* i *and* j*)* $\Leftrightarrow$ ξ_i *and* ξ_j *are independent.*

Definition 2 *(Fractional Covers (Janson, 2004)) Let* $\Gamma = (V, E)$ *be an undirected graph with* $V = [K]$.

- $C \subseteq V$ *is independent if the vertices in* C *are independent (i.e., no two vertices in* C *are connected).*
- $\mathbf{C} = \{C_j\}_{j=1}^J$*, with* $C_j \subseteq V$*, is a proper cover of* V *if each* C_j *is independent and* $\bigcup_{j=1}^J C_j = V$.
- $\mathbf{C} = \{(C_j, w_j)\}_{j=1}^J$*, with* $C_j \subseteq V$ *and* $w_j \in [0, 1]$*, is a proper exact fractional cover of* V *if* C_j *is independent and* $\forall i \in V$, $\sum_{j=1}^J w_j \mathbf{1}_{i \in C_j} = 1$*;* $\mathbf{w}(\mathbf{C}) = \sum_{j=1}^J w_j$ *is defined as the chromatic weight of* $\mathbf{C}$.
- *The fractional chromatic number* $\chi^*(\Gamma)$ *is the minimum chromatic weight over all proper exact fractional covers of the dependence graph* $\Gamma = (V, E)$.

Then we can obtain a chromatic PAC-Bayesian kl-bound for dependent random variables $\mathcal{S} = \{\xi_1, \ldots, \xi_K\}$.

Theorem 5 *In the same setting of Theorem 2 with the only difference that* $\mathcal{S} = \{\xi_k\}_{k=1}^K$ *is a sequence of dependent random variables. Let* $\chi^*(\mathcal{S})$ *denote the fractional chromatic number of the dependence graph of* $\mathcal{S}$*. Then with probability with at least* $1 - \delta$ *over the draw of* $\mathcal{S}$*, the following holds for any measure* ρ *over* $\mathcal{F}$*:*

$$\mathrm{kl}(\rho(r), \rho(R)) \leq \frac{\chi^*(\mathcal{S})}{K}[\mathcal{K}(\rho, \pi) + \ln(\frac{2}{\delta}\sqrt{\frac{K}{\chi^*(\mathcal{S})}})].$$

A previous result to deal with non-identically non-independently distributed data is the PAC-Bayesian chromatic bound in (Ralaivola et al., 2010) [Theorem 28], whose order is about $O(\sqrt{\frac{\ln K}{K}})$. The main difference between Ralaivola's bound and ours is as follow: the bound in Ralaivola's Theorem 28 is obtained by directly using a chromatic concentration inequality from (Janson, 2004) to bound the moment generating function (MGF) of the kl-divergence of dependent samples; instead, we employ graph decomposition techniques from (Janson, 2004) to encode dependent samples into independent sets and then apply our Lemma 1 to bound the MGF of the kl-divergence of independent samples, leading to a tighter chromatic PAC-Bayes bound of $O(\frac{\ln K}{K})$ in Theorem 5. Finally, applying the above theorem to bound $\mathrm{kl}(er(\mathcal{Q}), \widetilde{er}(\mathcal{Q}))$ and $\mathrm{kl}(\widetilde{er}(\mathcal{Q}), \widehat{er}(\mathcal{Q}))$ respectively, we obtain our chromatic PAC-Bayesian bound for meta-learning with dependent tasks and dependent samples.

Theorem 6 *For any given hyper-prior* $\mathcal{P}$*, with probability at least* $1 - \delta$ *over the draw of the training sample* $\{S_i\}_{i=1}^n$*, the following holds for any hyper-posterior* $\mathcal{Q}$*:*

$$er(\mathcal{Q}) \leq \widehat{er}(\mathcal{Q}) + \sqrt{\frac{\Delta_1}{2n}} + \sqrt{\frac{2\Delta_2 \widehat{er}(\mathcal{Q})}{mn}} + \frac{2\Delta_2}{mn},$$

where $\Delta_1 = \chi^*(\mathbf{D})[\mathcal{K}(\mathcal{Q}, \mathcal{P}) + \ln(\frac{2}{\delta}\sqrt{\frac{n}{\chi^*(\mathbf{D})}})]$, $\Delta_2 = \chi^*(\mathbf{S})\big[\mathcal{K}(\mathcal{Q}, \mathcal{P}) + \mathbf{E}_{P\sim\mathcal{Q}} \sum_{i=1}^n \mathcal{K}(Q_i, P) + \ln \frac{2\sqrt{mn}}{\delta\sqrt{\chi^*(\mathbf{S})}}\big]$, $\chi^*(\mathbf{D}), \chi^*(\mathbf{S})$ *denote the fractional chromatic numbers of the dependence graphs of* $\mathbf{D} = \{D_i\}_{i=1}^n$, $\mathbf{S} = \{S_i\}_{i=1}^n$.

It is not difficult to see that, when all samples $\mathbf{S} = \{z_{ij}\}_{i,j=1}^{n,m}$ in meta-learning are sampled independently, the proper exact fractional cover of $\Gamma(\mathbf{S}) = \{V, E\}$ is $\{(V, 1)\}$. Therefore $\chi^*(\mathbf{S}) = 1$. We can also obtain that $\chi^*(\mathbf{D}) = 1$ when all distributions $\{D_i\}_{i=1}^n$ are sampled independently from the task environment τ. In this case, Theorem 6 degrades to the PAC-Bayesian kl-bound for meta-learning with independence assumption in Theorem 3. Another example for calculating $\chi^*(\mathbf{D})$ in dependent meta learning setting is provided in Example 1 in Appendix D.

We should point out that in Theorem 6 (or Proposition 8 in the Appendix D), our environment-level complexity $\sqrt{\frac{\chi^*(\mathbf{D})}{2n}[\mathcal{K}(\mathcal{Q}, \mathcal{P}) + \ln(\frac{2}{\delta}\sqrt{\frac{n}{\chi^*(\mathbf{D})}})]}$ on $er(\mathcal{Q}) - \widetilde{er}(\mathcal{Q})$ is

Table 2. Comparisons of different PAC-Bayesian meta-learning methods. The average test bounds and test errors are reported over 20 test tasks (the $\pm$ shows the 95% confidence interval) in three different pixel-shuffled environments.

	100 Pixels Swaps		200 Pixels Swaps		300 Pixels Swaps	
Method	**Test Bound**	**Test Error** (%)	**Test Bound**	**Test Error** (%)	**Test Bound**	**Test Error** (%)
Variational Bayes	N/A	1.606 ± 0.001	N/A	1.962 ± 0.001	N/A	2.649 ± 0.130
MAML	N/A	1.876 ± 0.001	N/A	2.241 ± 0.002	N/A	2.788 ± 0.102
(Seeger, 2002)	0.133 ± 0.034	1.629 ± 0.000	0.285 ± 0.049	1.972 ± 0.001	0.408 ± 0.062	2.523 ± 0.001
(Pentina & Lampert, 2014)	0.190 ± 0.022	1.939 ± 0.001	0.240 ± 0.030	2.631 ± 0.002	0.334 ± 0.036	3.767 ± 0.003
(Amit & Meir, 2018)	0.126 ± 0.012	1.587 ± 0.001	0.197 ± 0.019	1.948 ± 0.001	0.270 ± 0.018	2.630 ± 0.001
(Rothfuss et al., 2021)	0.174 ± 0.023	1.921 ± 0.001	0.224 ± 0.030	2.634 ± 0.001	0.318 ± 0.036	3.754 ± 0.003
kl-bound (ours)	0.119 ± 0.024	1.746 ± 0.001	0.189 ± 0.027	2.594 ± 0.001	0.359 ± 0.042	2.993 ± 0.002
Catoni-bound (ours)	$\mathbf{0.093 \pm 0.027}$	$\mathbf{1.545 \pm 0.001}$	$\mathbf{0.128 \pm 0.025}$	$\mathbf{1.889 \pm 0.001}$	$\mathbf{0.210 \pm 0.035}$	$\mathbf{2.433 \pm 0.001}$

tighter than that in (Pentina & Lampert, 2015) [Theorem 3], which only bounds the deviation $er(\mathcal{Q}) - \widetilde{er}(\mathcal{Q})$ with $\sqrt{\frac{\chi^*(\mathbf{D})}{n}}[\mathcal{K}(\mathcal{Q}, \mathcal{P}) + \ln(J/\delta)]$, where J denotes the size of the dependence graph $\Gamma(\mathbf{D})$. Nevertheless, we also need to admit that in the current version we are unable to extend our PAC-Bayes Catoni-bound to the generalized meta-learning setting. This may somewhat imply that the kl-bound is more flexible than the Catoni-bound to be applied to learning settings where training data show some dependencies.

5. Experiments

In this section, we empirically demonstrate the effectiveness of our theoretical analysis for meta-learning over the classification and regression problems with deep neural networks. For classification problems, we directly set our kl-bound (Theorem 3) and Catoni-bound (Theorem 4) as minimization objectives, with the bounded cross-entropy loss (Pérez-Ortiz et al., 2021) for model optimization. For regression problems, we develop a Gibbs optimal hyper-posterior algorithm with Bayesian neural networks (GOHP-NN). Concretely, we choose the classical statistical inference method, called Stein Variational Gradient Descent (SVGD) (Liu & Wang, 2016), to approximate the Gibbs optimal hyper-posterior $\mathcal{Q}^*$ in Corollary 2 (with $C_1 = 2$). The mean squared error is selected as the loss function for regression problems. In practice, the squared loss is always bounded, so we choose it for model optimization for fair comparisons with existing methods. The detailed pseudo-code of our proposed meta-learning algorithms for classification and regression problems can be found in the Appendices E-F.

5.1. Experimental Setup

Classification Environments. We conduct classification experiments in three different task environments, based on the augmentations of the MNIST dataset (Yann, 1998). Each task from the same environment is constructed by a fixed number of pixel swaps to ensure the task relatedness. The three environments are created by swapping 100/200/300 pixels respectively to increase the classification difficulty. During the meta-training phase, we choose 10 training tasks, each of which is composed of 60,000 training examples; while in the meta-test phase, each task is constructed with fewer training samples (2,000). We choose a fully-connected network with 3 hidden layers and a linear output layer as backbone. All experiment details are set the same as that in (Amit & Meir, 2018). We report the test bounds and test errors on the novel tasks of various methods in Table 2.

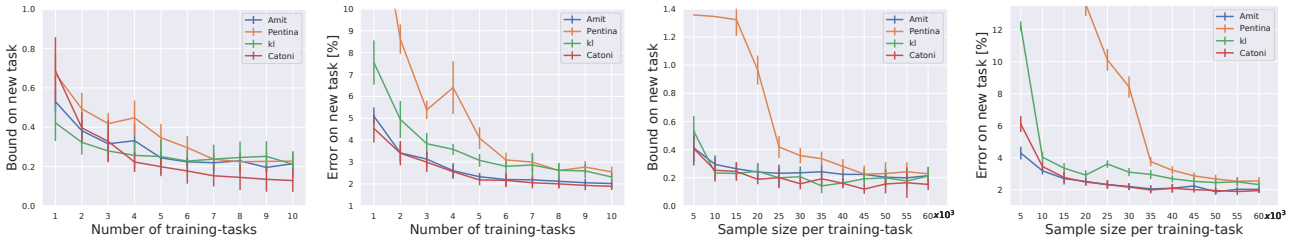
Regression Environments. We conduct regression experiments with one synthetic and four real-world meta-learning environments. The first synthetic environment is composed of regression tasks that can be interpreted as a 2-dimensional mixture of *Cauchy* distributions plus a random Gaussian Processes function. For the second environment, we employ datasets corresponding to different calibration sessions of *Swiss Free Electron Laser*(*SwissFEL*) (Milne et al., 2017). The other two environments are constructed by using the datasets from *PhysioNet* 2012 challenge (Silva et al., 2012), which contains the time series of electronic health measurements from patients, in terms of the Glasgow Coma Scale (*GCS*) and the hematocrit value (*HCT*). Finally, we create the *Berkeley-Sensor* environment where the tasks need to make prediction of temperature measurements corresponding to the sensors installed in different places of one building (Madden, 2004). More detailed information about the tasks in regression environments can be found in the Appendix F.1. We set the same experiment setup as that in (Rothfuss et al., 2021). Then we report the root mean squared errors (RMSE) over the novel tasks in Table 3. The results of other methods are directly cited from (Rothfuss et al., 2021) [Table 1]. Note that Rothfuss et al. do not report test bounds over the novel tasks in their original work. So Table 3 only reports the test errors for fair comparison.

5.2. Experimental Results

Classification Results. From Table 2, we can see that minimizing our proposed two generalization bounds can achieve competitive results w.r.t. existing PAC-Bayesian meta-learning methods, in terms of test bounds and test errors over the novel tasks. Particularly, our Catoni-bound method can obtain the tightest test bounds and lowest predic-

Table 3. Comparison of meta-learning algorithms in terms of test RMSE in 5 regression environments. Reported are mean and standard deviation across 5 seeds. Our GOHP-NN achieves competitive averaged error over 5 environments.

Method	Cauchy	SwissFel	Physionet-GCS	Physionet-HCT	Berkeley-Sensor
Vanilla BNN (Liu & Wang, 2016)	0.327 ± 0.008	0.529 ± 0.022	2.664 ± 0.274	3.938 ± 0.869	0.109 ± 0.004
MLL-GP (Fortuin & Rätsch, 2019)	0.216 ± 0.003	0.974 ± 0.093	1.654 ± 0.094	2.634 ± 0.144	0.058 ± 0.002
MLAP (Amit & Meir, 2018)	0.219 ± 0.004	0.486 ± 0.026	2.009 ± 0.248	2.470 ± 0.039	0.050 ± 0.005
MAML (Finn et al., 2017)	0.219 ± 0.004	0.730 ± 0.057	1.895 ± 0.141	2.413 ± 0.113	0.045 ± 0.003
BMAML (Yoon et al., 2018)	0.225 ± 0.004	0.577 ± 0.044	1.894 ± 0.062	2.500 ± 0.002	0.073 ± 0.014
PACOH-GP (Rothfuss et al., 2021)	0.209 ± 0.008	0.376 ± 0.024	$\mathbf{1.498 \pm 0.081}$	$\mathbf{2.361 \pm 0.047}$	0.065 ± 0.005
PACOH-NN (Rothfuss et al., 2021)	$\mathbf{0.195 \pm 0.001}$	0.372 ± 0.002	1.561 ± 0.061	2.405 ± 0.017	$\mathbf{0.043 \pm 0.001}$
GOHP-NN (ours)	0.198 ± 0.016	$\mathbf{0.333 \pm 0.013}$	1.521 ± 0.067	2.422 ± 0.013	$\mathbf{0.043 \pm 0.004}$



(a) (b) (c) (d)

Figure 1. Comparisons between our bounds (i.e., kl-bound & Catoni-bound) and other bounds (i.e., Pentina-bound & Amit-bound). Both test bounds and test errors are averaged over 20 meta-test classification tasks from 200-pixel-shuffled environment. (a)-(b): Results across a range of number n of training tasks. (c)-(d): Results across a range of sample size m per training task.

tion errors over all tasks from different environments. This is consistent with our theoretical results that the Catoni-bound is so far the tightest PAC-Bayesian bound for meta-learning. Meanwhile, the prediction performance of different methods gets worse with the increase of the number of pixel swaps. This indicates the importance of the task relatedness of environment to the success of meta-learning approaches.

Regression Results. From Table 3, we can observe that our proposed GOPH-NN algorithm can achieve comparable results w.r.t. the state-of-the-art PAC-Bayesian meta-learning method PACOH (Rothfuss et al., 2021) over 5 regression environments. Specifically, our GOPH-NN can obtain the lowest test errors on 2 regression environments, yield competitive test errors on other 3 regression environments. The detailed reasons that GOHP can achieve analogous results with PACOH can be found in Remark 2 in Appendix F. Therefore, we can conclude that our proposed PAC-Bayesian Catoni-bound, can obtain comparable performance with respect to the latest meta-learning regression methods, in terms of average test errors on the novel tasks.

Convergence Analysis. We compare the convergence rates of our two PAC-Bayesian bounds and other two classical bounds (Pentina & Lampert, 2014; Amit & Meir, 2018) in the 200-pixel-shuffled classification environment. The average test bounds and test errors are calculated over the novel tasks across a wide range of the number n of training tasks and across a large range of the sample size m per task. The visualization is shown in Figure 1. We can find that: (1) Both the test bounds and the test errors of all methods decrease with the increase of n and m. (2) Our proposed two bounds can achieve competitive performance with respect to the existing methods. Particularly, our PAC-Bayesian Catoni-bound obtains the tightest test bounds and the lowest test errors over the novel tasks, empirically validating the effectiveness of our theoretical analysis for meta-learning.

6. Conclusions

This work provides a unified demonstration framework of the PAC-Bayesian bounds for single-task learning and meta-learning. The tightest PAC-Bayesian kl-bound and Catoni-bound in single-task learning are generalized to the meta-learning setting, followed by two bound-minimizing meta-learning classification algorithms. Next, we show how to obtain the closed-form formula of the Gibbs optimal hyper-posterior by minimizing our Catoni-bound, leading to an efficient meta-learning regression algorithm. In addition, we obtain a chromatic PAC-Bayesian kl-bound for the more challenging meta-learning setting where training data show some dependencies. Experiments on classification and regression problems further validate the effectiveness of our proposed PAC-Bayesian bounds. In particular, our Catoni-bound obtains the tightest test bounds and the lowest test errors in classification problems, and achieves comparable results with existing methods in regression problems. Overall, we show how to derive two fast-rate PAC-Bayesian bounds for meta-learning, and show how to apply these bounds to different settings to yield more theoretical results.

Acknowledgements

Jiechao would like to thank Prof. Yong Liu from GSAI for insightful comments on the writing of this paper, and thank Prof. Zongfei Fu and Dr. Yin Gao from School of Mathematics in RUC for helpful discussions. We also thank all reviewers for their constructive suggestions to improve the quality of this paper. This work was supported in part by National Natural Science Foundation of China (61976220 and 61832017), Beijing Outstanding Young Scientist Program (BJJWZYJH012019100020098), and Large-Scale Pre-Training Program 468 of Beijing Academy of Artificial Intelligence (BAAI). Prof. Zhiwu Lu is the corresponding author of this paper.

References

Alquier, P. and Guedj, B. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018.

Amit, R. and Meir, R. Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *International Conference on Machine Learning (ICML)*, pp. 205–214, 2018.

Audibert, J.-Y. PAC-Bayesian aggregation and multi-armed bandits. *arXiv preprint arXiv:1011.3396*, 2010.

Baxter, J. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

Berend, D. and Tassa, T. Efficient bounds on bell numbers and on moments of sums of random variables. *Probability and Mathematical Statistics*, 30:185–205, 2010.

Catoni, O. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Institute of Mathematical Statistics, 2007.

Chen, J., Wu, X., Li, Y., LI, Q., Zhan, L., and Chung, F. A closer look at the training strategy for modern meta-learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 396–406, 2020.

Chen, Y., Guo, X., Wang, C., Qiu, J., Qi, G., Wang, M., and Li, H. Leveraging table content for zero-shot text-to-SQL with meta-learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3992–4000, 2021.

Denevi, G., Ciliberto, C., Stamos, D., and Pontil, M. Learning to learn around A common mean. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 10190–10200, 2018.

Farid, A. and Majumdar, A. Generalization bounds for meta-learning via PAC-Bayes and uniform stability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*, pp. 1126–1135, 2017.

Fortuin, V. and Rätsch, G. Deep Mean Functions for Meta-Learning in Gaussian Processes. *arXiv preprint arXiv:1901.08098*, 2019.

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. PAC-Bayesian learning of linear classifiers. In *International Conference on Machine Learning (ICML)*, pp. 353–360, 2009.

Haddouche, M., Guedj, B., Rivasplata, O., and Shawe-Taylor, J. PAC-Bayes unleashed: Generalisation bounds with unbounded losses. *Entropy*, 23(10), 2021.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Holland, M. J. PAC-Bayes under potentially heavy tails. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2711–2720, 2019.

Janson, S. Large deviations for sums of partly dependent random variables. *Random Structures & Algorithms*, 24 (3):234–248, 2004.

Li, K., Gupta, A., Reddy, A., Pong, V. H., Zhou, A., Yu, J., and Levine, S. MURAL: meta-learning uncertainty-aware rewards for outcome-driven reinforcement learning. In *International Conference on Machine Learning (ICML)*, pp. 6346–6356, 2021.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2370–2378, 2016.

Madden, S. Intel lab data. `http://db.csail.mit.edu/labdata/labdata.html`, 2004. Accessed: Sep 8, 2020.

Maurer, A. A note on the PAC-Bayesian theorem. *arXiv preprint arXiv:cs/0411099*, 2004.

Maurer, A. Algorithmic stability and meta-learning. *Journal of Machine Learning Research (JMLR)*, 6:967–994, 2005.

Maurer, A. Transfer bounds for linear feature learning. *Machine Learning*, 75(3):327–350, 2009.

Maurer, A., Pontil, M., and Romera-Paredes, B. The benefit of multitask representation learning. *Journal of Machine Learning Research (JMLR)*, 17:81:1–81:32, 2016.

McAllester, D. A. Some PAC-Bayesian theorems. In *Annual Conference on Learning Theory (COLT)*, pp. 230–234, 1998.

McAllester, D. A. Simplified PAC-Bayesian margin bounds. In *Annual Conference on Learning Theory (COLT)*, pp. 203–215, 2003.

Milne, C. J., Schietinger, T., Aiba, M., Alarcon, A., Alex, J., Anghel, A., Arsov, V., Beard, C., Beaud, P., Bettoni, S., et al. SwissFEL: the Swiss X-ray free electron laser. *Applied Sciences*, 7:720:1–720:57, 2017. URL https://www.mdpi.com/2076-3417/7/7/720.

Pentina, A. and Lampert, C. H. A PAC-Bayesian bound for lifelong learning. In *International Conference on Machine Learning (ICML)*, pp. 991–999, 2014.

Pentina, A. and Lampert, C. H. Lifelong learning with non-i.i.d. tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1540–1548, 2015.

Pérez-Ortiz, M., Rivasplata, O., Shawe-Taylor, J., and Szepesvári, C. Tighter risk certificates for neural networks. *Journal of Machine Learning Research (JMLR)*, 22(227):1–40, 2021.

Ralaivola, L., Szafranski, M., and Stempfel, G. Chromatic PAC-Bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes. *Journal of Machine Learning Research (JMLR)*, 11:1927–1956, 2010.

Rothfuss, J., Fortuin, V., Josifoski, M., and Krause, A. PACOH: Bayes-optimal meta-learning with PAC-Guarantees. In *International Conference on Machine Learning (ICML)*, pp. 9116–9126, 2021.

Seeger, M. W. PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of Machine Learning Research (JMLR)*, 3:233–269, 2002.

Seldin, Y., Laviolette, F., Cesa-Bianchi, N., Shawe-Taylor, J., and Auer, P. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12): 7086–7093, 2012.

Silva, I., Moody, G., Scott, D. J., Celi, L. A., and Mark, R. G. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *Computing in Cardiology*, 2012.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4077–4087, 2017.

Thrun, S. and Pratt, L. *Learning to Learn*. Kluwer Academic Publishers, 1998.

Wu, Q., Lin, Z., Wang, G., Chen, H., Karlsson, B. F., Huang, B., and Lin, C. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 9274–9281, 2020.

Yann, L. The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/, 1998.

Ye, H., Hu, H., Zhan, D., and Sha, F. Few-shot learning via embedding adaptation with set-to-set functions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8805–8814, 2020.

Yoon, J., Kim, T., Dia, O., Kim, S., Bengio, Y., and Ahn, S. Bayesian model-agnostic meta-learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 7343–7353, 2018.

Zhang, T. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

APPENDIX

A. Auxiliary Results

We first recall Lemma 2-5 before obtaining our results.

Lemma 2 *(Berend & Tassa, 2010)[Lemma 5.2] Let $\{\xi_k\}_{k=1}^K$ be independent random variables with $P(0 \leq \xi_k \leq a_k) = 1$. For $1 \leq k \leq K$, let η_k be a random variable assuming only the values 0 and a_k and having the same expectation as ξ_k, i.e. $\mathbf{E}\eta_k = \mathbf{E}\xi_k$. Then for any convex function $g : \mathbb{R} \to \mathbb{R}$,*

$$\mathbf{E}g(\sum_{k=1}^K \xi_k) \leq \mathbf{E}g(\sum_{k=1}^K \eta_k).$$

Lemma 3 *(Berend & Tassa, 2010)[Proposition 3.1] Let $\{\xi_k\}_{k=1}^K$ be a sequence of independent random variables with $P(0 \leq \xi_k \leq 1) = 1$, $\{\eta_k\}_{k=1}^K$ be a sequence of i.i.d. Bernoulli random variables with $\mathbf{E}\eta_k = K^{-1}(\sum_{k=1}^K \mathbf{E}\xi_k)$. Then for any convex function g,*

$$\mathbf{E}g(\sum_{k=1}^K \xi_k) \leq \mathbf{E}g(\sum_{k=1}^K \eta_k).$$

Remark 1 *Using Jensen's inequality of convex function g, we have $g(\sum_{k=1}^K \mathbf{E}\eta_k) = g(\sum_{k=1}^K \mathbf{E}\xi_k) \leq \mathbf{E}g(\sum_{k=1}^K \xi_k)$. Thus, what Lemma 3 does is to 'plug' the term $\mathbf{E}g(\sum_{k=1}^K \xi_k)$ into the Jensen's inequality $g(\sum_{k=1}^K \mathbf{E}\eta_k) \leq \mathbf{E}g(\sum_{k=1}^K \eta_k)$. So the inequality in Lemma 3 is truly tight.*

Corollary 3 *(Lemma 1 in the main paper) In the setting of Lemma 2 or Lemma 3, for any convex function g, $\mathbf{E}g(\frac{1}{K}\sum_{k=1}^K \xi_k) \leq \mathbf{E}g(\frac{1}{K}\sum_{k=1}^K \eta_k)$.*

Proof. The composition function $(g \circ f)(x)$ of the convex function g and the linear function $f(x) = ax + b$ is still convex. Setting $a = \frac{1}{K}, b = 0$ completes the proof. ■

Lemma 4 *(Change of Measure) Let $\mathcal{F}$ be a set of random variables f. Let $\mathcal{S} = \{\xi_k\}_{k=1}^K$ be a sequence of random variables with each component ξ_k ($k \in [K]$) drawn independently according to the measure μ_k over the set A_k. Then, for any functions $R(f)$, $r(f)$ over $\mathcal{F}$, either of which may be a statistic of $\mathcal{S}$, any reference measure π over $\mathcal{F}$, any $\lambda > 0$, and any convex function $\mathcal{D} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, the following holds for any measure ρ over $\mathcal{F}$:*

$$\mathcal{D}(\mathbf{E}_\rho R(f), \mathbf{E}_\rho r(f))) \leq \frac{\mathcal{K}(\rho, \pi) + \ln \mathbf{E}_{f\sim\pi} \mathrm{e}^{\lambda \mathcal{D}(R(f), r(f))}}{\lambda},$$

where $\mathcal{K}(\rho, \pi)$ denotes the KL-divergence between the distributions ρ and π.

Lemma 5 *(Catoni, 2007)[Lemma 1.1.1] Given independent input-output pairs $\{(x_k, y_k)\}_{k=1}^K \in (\mathcal{X} \times \mathcal{Y})^K$ and a class of classification rules $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$, let $R(f) = K^{-1}\sum_{k=1}^K P[f(x_k) \neq y_k]$, $r(f) = K^{-1}\sum_{k=1}^K \mathbf{1}[f(x_k) \neq y_k]$. Then for any real constant $\lambda \in \mathbb{R}$, and any $f \in \mathcal{F}$,*

$$\mathbf{E}\exp\{\lambda[\Phi_{\frac{\lambda}{K}}(R(f)) - r(f)]\} \leq 1,$$

where [2] *$\Phi_a(p) = -a^{-1}\ln\{1 - [1 - \exp(-a)]p\}$, $a \in \mathbb{R}$, $p \in [0, 1]$.*

B. Proofs of the PAC-Bayesian Bounds for Meta-Learning with Independent Samples

B.1. Proof of PAC-Bayesian kl-Bound

Proposition 1 *(Part of Theorem 2 in the main paper) In the setting of Lemma 4, set $\mathcal{D}(q, p) = \mathrm{kl}(p, q)$, where $\mathrm{kl}(p, q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}$. Let $R(f) = \frac{1}{K}\sum_{k=1}^K \mathbf{E}_{\xi_k} g_k(f, \xi_k)$, $r(f) = \frac{1}{K}\sum_{k=1}^K g_k(f, \xi_k)$, where $g_k : \mathcal{F} \times A_k \to [0, 1]$ is a bounded function. Then with probability at least $1 - \delta$ over the draw of $\mathcal{S}$, the following holds for any measure ρ:*

$$\mathrm{kl}(\rho(r), \rho(R)) \leq \frac{\mathcal{K}(\rho, \pi) + \ln\frac{2\sqrt{K}}{\delta}}{K}.$$

In particular, we have the explicit generalization bound:

$$|\rho(R) - \rho(r)| \leq \sqrt{\frac{2\Delta\rho(r)}{K}} + \frac{2\Delta}{K}.$$

where $\Delta = \mathcal{K}(\rho, \pi) + \ln\frac{2\sqrt{K}}{\delta}$.

Proof. For any fixed $f \in \mathcal{F}$, let $\{\eta_k\}_{k=1}^K$ be i.i.d. Bernoulli random variables with $\mathbf{E}\eta_k = \frac{1}{K}\sum_{k=1}^K \mathbf{E}_{\xi_k} g_k(f, \xi_k)$. Note that $\mathrm{kl}(p, q)$ is a convex function with respect to p and exp is a nondecreasing convex function, hence $\exp\{\lambda\,\mathrm{kl}(p, q)\}$ is a convex function with respect to p ($\lambda > 0$). Then setting $\lambda = K$, we have

$$\begin{aligned}
&\mathbf{E}_{\mathcal{S}}\mathrm{e}^{K\,\mathrm{kl}(r(f), R(f))} \\
=&\mathbf{E}\mathrm{e}^{K\,\mathrm{kl}(\frac{1}{K}\sum_{k=1}^K g_k(f,\xi_k), \frac{1}{K}\sum_{k=1}^K \mathbf{E}\eta_k)} \\
\leq&\mathbf{E}\mathrm{e}^{K\,\mathrm{kl}(\frac{1}{K}\sum_{k=1}^K \eta_k, \frac{1}{K}\sum_{k=1}^K \mathbf{E}\eta_k)} \quad \text{(Corollary 3)} \\
\leq&2\sqrt{K}.
\end{aligned}$$

The last inequality holds due to the fact that for a binomial random variable $\eta \sim B(K, \mu)$, $\mathbf{E}\mathrm{e}^{K\,\mathrm{kl}(\frac{\eta}{K}, \mu)} = \sum_{k=0}^K \binom{K}{k}(\frac{k}{K})^k(\frac{K-k}{K})^{K-k} \in [\sqrt{K}, 2\sqrt{K}]$ (Maurer, 2004). Then recalling Lemma 4 and Markov's inequality

[2] $\Phi_a(p)$ is a one-to-one increasing function with respect to $p \in [0, 1]$, and is convex when $a > 0$.

we have

$$\begin{aligned}
&\mathrm{kl}(\rho(r),\rho(R)) \le \frac{1}{K}\left[\mathcal{K}(\rho,\pi) + \ln \mathbf{E}_{f\sim\pi} \mathrm{e}^{K\,\mathrm{kl}(r(f),R(f))}\right] \\
\le& \frac{\mathcal{K}(\rho,\pi)}{K} + \frac{1}{K}\ln \mathbf{E}_{\mathcal{S}}\mathbf{E}_{f\sim\pi}\mathrm{e}^{K\,\mathrm{kl}(r(f),R(f))}/\delta \\
=& \frac{\mathcal{K}(\rho,\pi)}{K} + \frac{1}{K}\ln \mathbf{E}_{f\sim\pi}\mathbf{E}_{\mathcal{S}}\mathrm{e}^{K\,\mathrm{kl}(r(f),R(f))}/\delta \quad \text{(Fubini)} \\
\le& \frac{\mathcal{K}(\rho,\pi)}{K} + \frac{1}{K}\ln(2\sqrt{K}/\delta),
\end{aligned}$$

which completes the proof of the first assertion. Using Pinsker's inequality $\mathrm{kl}(p,q) \ge 2(p-q)^2$ we can directly obtain an explicit upper bound on $|\rho(R)-\rho(r)|$ with probability at least $1-\delta$ as follow:

$$|\rho(R)-\rho(r)| \le \sqrt{\frac{\mathcal{K}(\rho,\pi)+\ln\frac{2\sqrt{K}}{\delta}}{2K}}.$$

For the second assertion, note that the first assertion is equivalent to the below statement (McAllester, 2003) [Eq.(5)]:

$$\forall\rho, \rho(R) \le \sup\left\{\epsilon : \mathrm{kl}(\rho(r),\epsilon) \le \frac{\mathcal{K}(\rho,\pi)+\ln\frac{2\sqrt{K}}{\delta}}{K}\right\}.$$

Thus we can use the tighter version of Pinsker's inequality $\mathrm{kl}(p,q) \ge \frac{(p-q)^2}{2q}$, when $q > p$. Then, if $\mathrm{kl}(p,q) \le x$, we have $[q-(p+x)]^2 - x^2 - 2px \le 0$ and so $|q-p| \le x + \sqrt{x^2+2px} \le 2x + \sqrt{2px}$. Combining this with the bound on $\mathrm{kl}(\rho(R),\rho(r))$ in the first assertion, we can give a high-probability bound on $|\rho(R)-\rho(r)|$ with

$$\sqrt{\frac{2\rho(r)[\mathcal{K}(\rho,\pi)+\ln\frac{2\sqrt{K}}{\delta}]}{K}} + \frac{2[\mathcal{K}(\rho,\pi)+\ln\frac{2\sqrt{K}}{\delta}]}{K}. \quad \blacksquare$$

We use the above proposition to bound $\mathrm{kl}(er(\mathcal{Q}),\widetilde{er}(\mathcal{Q}))$ and $\mathrm{kl}(\widetilde{er}(\mathcal{Q}),\widehat{er}(\mathcal{Q}))$ respectively for meta-learning.

Proposition 2 *For any $\delta \in (0,1)$, with probability at least $1-\delta$ over the draw of n distributions $\{D_i\}_{i=1}^n$, the following holds for any hyper-posterior $\mathcal{Q}$:*

$$|er(\mathcal{Q}) - \widetilde{er}(\mathcal{Q})| \le \sqrt{\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})+\ln\frac{2\sqrt{n}}{\delta}}{2n}}.$$

Proof. Notice that

$$\begin{aligned}
er(\mathcal{Q}) &= \mathbf{E}_{P\sim\mathcal{Q}}\mathbf{E}_{(D,S)\sim\tau\times D^m} er(Q(S,P),D) \\
\widetilde{er}(\mathcal{Q}) &= \mathbf{E}_{P\sim\mathcal{Q}}\frac{1}{n}\sum_{i=1}^n er(Q(S_i,P),D_i).
\end{aligned}$$

Recalling Proposition 1, we set $K=n$, $f=P$, $\pi=\mathcal{P}$, $\rho=\mathcal{Q}$, $\xi_k=(D_i,S_i)$, $g_k(f,\xi_k) = \mathbf{E}_{h\sim Q(S_i,P)}\mathbf{E}_{z\sim D_i} l(h,z) \in [0,1]$. Thus with Pinsker's inequality $\mathrm{kl}(p,q) \ge 2(p-q)^2$,

$$|er(\mathcal{Q}) - \widetilde{er}(\mathcal{Q})| \le \sqrt{\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})+\ln\frac{2\sqrt{n}}{\delta}}{2n}}. \blacksquare$$

Proposition 3 *For any hyper-prior $\mathcal{P}$, any $\delta \in (0,1)$, with probability at least $1-\delta$ over the draw of the training sample $S=\{S_i\}_{i=1}^n$, the following holds for any hyper-posterior $\mathcal{Q}$:*

$$|\widetilde{er}(\mathcal{Q}) - \widehat{er}(\mathcal{Q})| \le \sqrt{\frac{2\Delta\widehat{er}(\mathcal{Q})}{mn}} + \frac{2\Delta}{mn},$$

where $\Delta = \mathcal{K}(\mathcal{Q},\mathcal{P}) + \mathbf{E}_{P\sim\mathcal{Q}}\sum_{i=1}^n \mathcal{K}(Q_i,P) + \ln\frac{2\sqrt{mn}}{\delta}$.

Proof. Notice that

$$\begin{aligned}
\widetilde{er}(\mathcal{Q}) &= \mathbf{E}_{P\sim\mathcal{Q}}\mathbf{E}_{(h_1,\cdots,h_n)\sim Q_1\times\cdots\times Q_n}\frac{1}{n}\sum_{i=1}^n \mathbf{E}_{z\sim D_i} l(h_i,z), \\
\widehat{er}(\mathcal{Q}) &= \mathbf{E}_{P\sim\mathcal{Q}}\mathbf{E}_{(h_1,\cdots,h_n)\sim Q_1\times\cdots\times Q_n}\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m l(h_i,z_{ij}).
\end{aligned}$$

Recall Proposition 1, we set $f=(P,h_1,...,h_n)$, $\pi=\mathcal{P}\times P^n$, $\rho=\mathcal{Q}\times\prod_{i=1}^n Q_i$, where $Q_i=Q(S_i,P)$, $\xi_k=z_{ij}$, $g_k(f,\xi_k)=l(h_i,z_{ij})$. With probability $\ge 1-\delta$ we have,

$$|\widetilde{er}(\mathcal{Q}) - \widehat{er}(\mathcal{Q})| \le \sqrt{\frac{2\Delta_1\widehat{er}(\mathcal{Q})}{mn}} + \frac{2\Delta_1}{mn}.$$

where $\Delta_1 = \mathcal{K}(\mathcal{Q}\times\prod_{i=1}^n Q_i, \mathcal{P}\times P^n) + \ln\frac{2\sqrt{mn}}{\delta}$. Further, notice that $\mathcal{K}(\mathcal{Q}\times\prod_i^n Q_i, \mathcal{P}\times P^n) = \mathbf{E}_{\mathcal{Q}\times\prod_{i=1}^n Q_i}\ln\frac{\mathrm{d}(\mathcal{Q}\times\prod_{i=1}^n Q_i)}{\mathrm{d}(\mathcal{P}\times P^n)} = \mathbf{E}_{\mathcal{Q}\times\prod_{i=1}^n Q_i}\left(\ln\frac{\mathrm{d}\mathcal{Q}}{\mathrm{d}\mathcal{P}} + \sum_{i=1}^n\ln\frac{\mathrm{d}Q_i}{\mathrm{d}P}\right) = \mathcal{K}(\mathcal{Q},\mathcal{P}) + \mathbf{E}_{P\sim\mathcal{Q}}\sum_{i=1}^n\mathcal{K}(Q_i,P)$, which completes the whole proof. $\blacksquare$

Proof of Theorem 3 in the main paper.
Note that the generalization bounds in Propositions 2-3 are both two-sided. Actually a one-sided version can be stated by replacing $2/\delta$ in the two-sided version with $1/\delta$. Therefore, combining the one-sided bounds in Propositions 2-3, and applying the union bound to these two high-probability inequalities finishes the proof. $\blacksquare$

B.2. Proof of PAC-Bayesian Catoni-Bound

Proposition 4 *(Part of Theorem 2 in the main paper) Set $\mathcal{D}(q,p) = \Phi_{\frac{\lambda}{K}}(q) - p$ in Lemma 4. Let $R(f) = \frac{1}{K}\sum_{k=1}^K \mathbf{E}_{\xi_k} g_k(f,\xi_k)$, $r(f) = \frac{1}{K}\sum_{k=1}^K g_k(f,\xi_k)$, where $g_k : \mathcal{F}\times A_k \to [0,1]$ is a bounded function. Then with probability at least $1-\delta$ over the draw of $\mathcal{S}$, the following*

holds for any measure ρ over $\mathcal{F}$:

$$\begin{aligned}\rho(R) \leq & \Phi_{\frac{\lambda}{K}}^{-1}[\rho(r)+\frac{\mathcal{K}(\rho,\pi)+\ln(1/\delta)}{\lambda}] \\ = & \frac{1-\exp\{-\frac{\lambda}{K}\rho(r)-\frac{\mathcal{K}(\rho,\pi)+\ln(1/\delta)}{K}\}}{1-\exp\{-\lambda/K\}} \\ \leq & \frac{\lambda\rho(r)}{K[1-\exp\{-\lambda/K\}]}+\frac{\mathcal{K}(\rho,\pi)+\ln(1/\delta)}{K[1-\exp\{-\lambda/K\}]}.\end{aligned}$$

Proof. We proceed with similar arguments as that in the proof of Proposition 1. Denote $\{\eta_k\}_{k=1}^K$ as independent Bernoulli random variables with $\mathbf{E}_{\xi_k} g_k(f,\xi_k) = \mathbf{E}\,\eta_k$. Since $\mathcal{D}(q,p)$ is a convex function of p, and exp function is a nondecreasing convex function, thus $\forall\lambda\in\mathbb{R}^+$, $\exp\{\lambda\mathcal{D}(q,p)\}$ is also a convex function with respect to p. Then we have

$$\begin{aligned}&\mathbf{E}_{\mathcal{S}}\mathrm{e}^{\lambda\mathcal{D}(R(f),r(f))} \\ =&\mathbf{E}_{\mathcal{S}}\mathrm{e}^{\lambda\mathcal{D}(\frac{1}{K}\sum_{k=1}^K \mathbf{E}_{\xi_k} g_k(f,\xi_k),\frac{1}{K}\sum_{k=1}^K g_k(f,\xi_k))} \\ =&\mathbf{E}\mathrm{e}^{\lambda\mathcal{D}(\frac{1}{K}\sum_k \mathbf{E}\eta_k,\frac{1}{K}\sum_{k=1}^K g_k(f,\xi_k))} \\ \leq&\mathbf{E}\mathrm{e}^{\lambda\mathcal{D}(\frac{1}{K}\sum_k \mathbf{E}\eta_k,\frac{1}{K}\sum_k \eta_k)} \quad \text{(Corollary 3)} \\ \leq&1 \quad \text{(Lemma 5)}.\end{aligned}$$

Recalling Lemma 4 and Markov's inequality we have

$$\begin{aligned}&\Phi_{\frac{\lambda}{K}}(\rho(R))-\rho(r)=\mathcal{D}(\rho(R),\rho(r)) \\ \leq&\frac{1}{\lambda}\big[\mathcal{K}(\rho,\pi)+\ln\mathbf{E}_{f\sim\pi}\mathrm{e}^{\lambda\mathcal{D}(R(f),r(f))}\big] \\ \leq&\frac{\mathcal{K}(\rho,\pi)}{\lambda}+\frac{1}{\lambda}\ln\mathbf{E}_{\mathcal{S}}\mathbf{E}_{f\sim\pi}\mathrm{e}^{\lambda\mathcal{D}(R(f),r(f))}/\delta \\ =&\frac{\mathcal{K}(\rho,\pi)+\ln(1/\delta)}{\lambda}+\frac{\ln\mathbf{E}_{f\sim\pi}\mathbf{E}_{\mathcal{S}}\mathrm{e}^{\lambda\mathcal{D}(R(f),r(f))}}{\lambda} \text{ (Fubini)} \\ \leq&\frac{\mathcal{K}(\rho,\pi)+\ln(1/\delta)}{\lambda}.\end{aligned}$$

Further, notice that the inverse function of $\Phi_a(p)$ is $\Phi_a^{-1}(q) = \frac{1-\exp\{-aq\}}{1-\exp\{-a\}}$, and the basic inequality $1-\exp(-x)\leq x$, we finish the whole proof. ■

We can derive a more concise corollary of the above result.

Corollary 4 *In the setting of Proposition 4, let $\lambda = K\ln C$, where $C>1$, then $\exp\{-\frac{\lambda}{K}\}=\frac{1}{C}$. Then with probability at least $1-\delta$ over the draw of $\mathcal{S}$, the following holds for any probability measure ρ:*

$$\rho(R)\leq\frac{C\ln C}{C-1}\rho(r)+\frac{C}{C-1}\frac{\mathcal{K}(\rho,\pi)+\ln(1/\delta)}{K}$$

With almost the same proceedings as that in the proof of Propositions 2-3, we can immediately yield the following two propositions for meta-learning with the use of Proposition 4 and Corollary 4. The detailed proof is thus omitted.

Proposition 5 *For any $\delta\in(0,1)$, with probability at least $1-\delta$ over the draw of n distributions $\{D_i\}_{i=1}^n$, the following holds for any $C>1$ and any hyper-posterior $\mathcal{Q}$:*

$$er(\mathcal{Q})\leq\frac{C\ln C}{C-1}\widetilde{er}(\mathcal{Q})+\frac{C}{C-1}\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})+\ln(1/\delta)}{n}.$$

Proposition 6 *For any hyper-prior $\mathcal{P}$, any $\delta\in(0,1)$, with probability at least $1-\delta$ over the draw of the training sample $\mathbf{S}=\{S_i\}_{i=1}^n$, the following holds for any hyper-posterior $\mathcal{Q}$:*

$$\begin{aligned}\widetilde{er}(\mathcal{Q})\leq&\frac{C\ln C}{C-1}\widehat{er}(\mathcal{Q})+ \\ &\frac{C}{C-1}\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\sum_{i=1}^n\mathcal{K}(Q_i,P)+\ln(1/\delta)}{nm}.\end{aligned}$$

Proof of Theorem 4 in the main paper. Combining Proposition 5 and Proposition 6, we have with probability at least $1-\delta$, $\forall C_1,C_2>1$,

$$\begin{aligned}er(\mathcal{Q})\leq&\frac{C_1C_2\ln C_1\ln C_2}{(C_1-1)(C_2-1)}\widehat{er}(\mathcal{Q})+\frac{C_1\big(\mathcal{K}(\mathcal{Q},\mathcal{P})+\ln(2/\delta)\big)}{n(C_1-1)} \\ &+\frac{C_1C_2\ln C_1\big(\mathcal{K}(\mathcal{Q},\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\sum_{i=1}^n\mathcal{K}(Q_i,P)+\ln(2/\delta)\big)}{(C_1-1)(C_2-1)nm}.\end{aligned} \tag{4}$$

C. Proof of the Theoretical Results of Gibbs Optimal Hyper-posterior

This section details the proof of the theoretical results about optimizing our PAC-Bayesian Catoni-bound with Gibbs optimal hyper-posterior. We first give the following helpful lemma that exhibits the explicit form of the Gibbs optimal posterior to certain optimization problems.

Lemma 6 *(Catoni, 2007)[Lemma 1.1.3] Let $\phi:\mathcal{F}\to\mathbb{R}$ be a measurable function. Then for any predefined probability measure $\pi\in\mathcal{M}_1(\mathcal{F})$, for any $\beta>0$, the minimizing probability measure ρ^* of the below optimization problem*

$$\min_{\rho\in\mathcal{M}_1(\mathcal{F})}\beta\mathbf{E}_{f\sim\rho}\phi(f)+\mathcal{K}(\rho,\pi)$$

has the following explicit form:

$$\frac{\mathrm{d}\rho^*}{\mathrm{d}\pi}(f)=\frac{\mathrm{e}^{-\beta\phi(f)}}{Z},$$

where $Z=\int_{\mathcal{F}}\mathrm{e}^{-\beta\phi(f)}\mathrm{d}\pi(f)$ is the normalization constant.

Proof of Corollary 1 in the main paper . Recall Theorem 4 in the main paper. We can actually rewrite the PAC-Bayesian meta-learning bound on $er(\mathcal{Q})$ in Theorem 4 as follow if we set $C_2 = \mathrm{e}$:

$$\underbrace{\frac{\mathrm{e}C_1 \ln C_1}{(C_1-1)(\mathrm{e}-1)}\widehat{er}(\mathcal{Q}) + \frac{\mathrm{e}C_1 \ln C_1 \mathbf{E}_{P\sim\mathcal{Q}} \sum_{i=1}^n \mathcal{K}(Q_i, P)}{nm(C_1-1)(\mathrm{e}-1)}}_{\mathbf{I}}$$

$$+\frac{C_1\big(\mathcal{K}(\mathcal{Q},\mathcal{P}) + \ln\frac{2}{\delta}\big)}{n(C_1-1)} + \frac{\mathrm{e}C_1 \ln C_1\big(\mathcal{K}(\mathcal{Q},\mathcal{P}) + \ln\frac{2}{\delta}\big)}{nm(C_1-1)(\mathrm{e}-1)}.$$

Actually, term $\mathbf{I}$ can be written as follow if we set Q_i^* as the Gibbs optimal posterior:

$$\frac{\mathrm{e}C_1 \ln C_1}{(C_1-1)(\mathrm{e}-1)}\big[\widehat{er}(\mathcal{Q}) + \frac{\mathbf{E}_{P\sim\mathcal{Q}} \sum_{i=1}^n \mathcal{K}(Q_i^*, P)}{nm}\big].$$

Then we have

$$\begin{aligned}
&\widehat{er}(\mathcal{Q}) + \frac{\mathbf{E}_{P\sim\mathcal{Q}} \sum_{i=1}^n \mathcal{K}(Q_i^*, P)}{nm}\\
=&\mathbf{E}_{P\sim\mathcal{Q}}\frac{1}{nm}\sum_{i=1}^n \big[m\widehat{er}(Q_i^*, S_i) + \mathcal{K}(Q_i^*, P)\big]\\
=&\mathbf{E}_{P\sim\mathcal{Q}}\frac{1}{nm}\sum_{i=1}^n \mathbf{E}_{h\sim Q_i^*}\big[m\widehat{er}(h, S_i) + \ln\frac{\mathrm{d}Q_i^*}{\mathrm{d}P}\big]\\
=&\mathbf{E}_{P\sim\mathcal{Q}}\frac{1}{nm}\sum_{i=1}^n \mathbf{E}_{h\sim Q_i^*}\big[m\widehat{er}(h, S_i) + \ln\frac{\mathrm{e}^{-m\widehat{er}(h,S_i)}}{Z(S_i,P)}\big]\\
=&\mathbf{E}_{P\sim\mathcal{Q}}\frac{1}{nm}\sum_{i=1}^n \big[-\ln Z(S_i,P)\big],
\end{aligned}$$

which finishes the whole proof. ■

Proof of Corollary 2 in the main paper. Recalling Corollary 1 we can obtain the form of the optimal Gibbs optimal hyper-posterior as follow:

$$\begin{aligned}
\mathcal{Q}^* =& \mathop{\arg\min}_{\mathcal{Q}\in\mathcal{M}_1(\mathcal{M}_1(\mathcal{H}))}\\
&\Big\{\frac{\mathrm{e}C_1 \ln C_1}{(C_1-1)(\mathrm{e}-1)}\mathbf{E}_{P\sim\mathcal{Q}}\frac{1}{nm}\sum_{i=1}^n[-\ln Z(S_i,P)]\\
&+\frac{C_1\big(\mathcal{K}(\mathcal{Q},\mathcal{P}) + \ln\frac{2}{\delta}\big)}{n(C_1-1)} + \frac{\mathrm{e}C_1 \ln C_1\big(\mathcal{K}(\mathcal{Q},\mathcal{P}) + \ln\frac{2}{\delta}\big)}{nm(C_1-1)(\mathrm{e}-1)}\Big\}\\
=& \mathop{\arg\min}_{\mathcal{Q}\in\mathcal{M}_1(\mathcal{M}_1(\mathcal{H}))}\\
&\Big\{\frac{\mathrm{e}C_1 \ln C_1}{(C_1-1)(\mathrm{e}-1)}\mathbf{E}_{P\sim\mathcal{Q}}\frac{1}{nm}\sum_{i=1}^n[-\ln Z(S_i,P)]\\
&+\big[\frac{C_1}{n(C_1-1)} + \frac{eC_1 \ln C_1}{nm(C_1-1)(\mathrm{e}-1)}\big]\mathcal{K}(\mathcal{Q},\mathcal{P})\Big\}\\
=& \mathop{\arg\min}_{\mathcal{Q}\in\mathcal{M}_1(\mathcal{M}_1(\mathcal{H}))}\Big[\mathbf{E}_{P\sim\mathcal{Q}}\frac{-\beta}{nm}\sum_{i=1}^n \ln Z(S_i,P) + \mathcal{K}(\mathcal{Q},\mathcal{P})\Big].
\end{aligned}$$

Applying Lemma 6 to the above minimization problem obtains the close-form of the Gibbs optimal hyper-posterior. ■

D. Proof of the PAC-Bayesian kl-Bound for Meta-Learning with Dependent Samples

We first give a fundamental lemma about the property of the exact proper fraction cover of the dependence graph.

Lemma 7 *(Janson, 2004)[Lemma 3.1] If* $\mathbf{C} = \{(C_j, w_j)\}_{j=1}^J$ *is an exact fractional cover of the dependence graph* $\Gamma = (V, E)$*, with* $V = [K]$*, then*

$$\forall \vec{t} \in \mathbb{R}^K, \quad \sum_{k=1}^K t_k = \sum_{j=1}^J w_j \sum_{k\in C_j} t_k.$$

Further, $K = \sum_{j=1}^J w_j |C_j|$*, where* $|C_j|$ *is the size of* C_j*.*

Proposition 7 *(Theorem 5 in the main paper) Let* $\mathcal{F}$ *be a set of random variables* f*. Let* $\mathcal{S} = (\xi_1, \cdots, \xi_K)$ *be a size-*K *random vector with each component* $\xi_k (k \in [K])$ *drawn according to the measure* μ_k *over the set* A_k*. Let* $R(f,\mathcal{S}) = \frac{1}{K}\sum_{k=1}^K \mathbf{E}_{\xi_k} g_k(f,\xi_k)$*,* $r(f,\mathcal{S}) = \frac{1}{K}\sum_{k=1}^K g_k(f,\xi_k)$*, where* $g_k : \mathcal{F}\times A_k \to [0,1]$ *is a bounded function. Then for any reference measure* π *over* $\mathcal{F}$*, with probability at least* $1-\delta$ *over the draw of* $\mathcal{S}$*, the following holds for any measure* ρ*:*

$$\begin{aligned}
&\mathrm{kl}(\mathbf{E}_\rho r(f,\mathcal{S}), \mathbf{E}_\rho R(f,\mathcal{S}))\\
&\le \frac{\chi^*(\mathcal{S})}{K}[\mathcal{K}(\rho,\pi) + \ln(\frac{2}{\delta}\sqrt{\frac{K}{\chi^*(\mathcal{S})}})],
\end{aligned}$$

where $\chi^*(\mathcal{S})$ *denotes the fractional chromatic number of the dependence graph of* $\mathcal{S}$*.*

Proof. According to Lemma 3.2 in (Janson, 2004), the fractional chromatic number is achieved when the cover is the exact proper fractional cover. Therefore, let us just consider $\mathbf{C} = \{(C_j, w_j)\}_{j=1}^J$ as the exact proper fractional cover of the dependence graph $\Gamma(\mathcal{S})$. Denote $\pi_j = \frac{w_j|C_j|}{K}$, we have $\sum_j \pi_j = 1$. Let $\mathcal{S}^{(j)} = \{\xi_k\}_{k\in C_j}$, hence the elements in $\mathcal{S}^{(j)}$ are all independent. Then we have

$$\begin{aligned}
&\sum_{j=1}^J \pi_j r(f,\mathcal{S}^{(j)}) = \sum_{j=1}^J \frac{w_j|C_j|}{K}\frac{1}{|C_j|}\sum_{l\in C_j} g_l(f,\xi_l)\\
=&\frac{1}{K}\sum_{j=1}^J w_j \sum_{l\in C_j} g_l(f,\xi_l)\\
=&\frac{1}{K}\sum_{k=1}^K g_k(f,\xi_k) = r(f,\mathcal{S}) \quad \text{(Lemma 7)}.
\end{aligned}$$

Meanwhile we also have

$$\sum_{j=1}^{J}\pi_j\mathbf{E}_{\mathcal{S}^{(j)}}r(f,\mathcal{S}^{(j)}) = \frac{1}{K}\sum_{j=1}^{J}w_j\sum_{l\in C_j}\mathbf{E}_{\xi_l}g_l(f,\xi_l)$$
$$=\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}_{\xi_k}g_k(f,\xi_k) = R(f,\mathcal{S}) \quad \text{(Lemma 7)}.$$

Then we have

$$\begin{aligned}
&\mathrm{kl}(\mathbf{E}_\rho r(f,\mathcal{S}),\mathbf{E}_\rho R(f,\mathcal{S}))\\
=&kl\big(\mathbf{E}_\rho\sum_{j=1}^{J}\pi_j r(f,\mathcal{S}^{(j)}),\mathbf{E}_\rho\sum_{j=1}^{J}\pi_j\mathbf{E}_{\mathcal{S}^{(j)}}r(f,\mathcal{S}^{(j)})\big)\\
\le&\sum_{j=1}^{J}\pi_j\mathbf{E}_\rho kl\big(r(f,\mathcal{S}^{(j)}),R(f,\mathcal{S}^{(j)})\big) \quad \text{(Jensen)}\\
\le&\sum_{j=1}^{J}\pi_j\frac{1}{|C_j|}\big[\mathcal{K}(\rho,\pi)+\ln\mathbf{E}_{f\sim\pi}\mathrm{e}^{|C_j|\,\mathrm{kl}(r(f,\mathcal{S}^{(j)}),R(f,\mathcal{S}^{(j)}))}\big]\\
\le&\sum_{j=1}^{J}\frac{\pi_j}{|C_j|}\big[\mathcal{K}(\rho,\pi)+\ln\frac{\mathbf{E}_{\mathcal{S}^{(j)}}\mathbf{E}_{f\sim\pi}\mathrm{e}^{|C_j|\,\mathrm{kl}(r(f,\mathcal{S}^{(j)}),R(f,\mathcal{S}^{(j)}))}}{\delta}\big]\\
\le&\sum_{j=1}^{J}\frac{\pi_j}{|C_j|}\big[\mathcal{K}(\rho,\pi)+\ln\frac{2\sqrt{|C_j|}}{\delta}\big]\\
=&\frac{\mathbf{w}\big(\mathcal{K}(\rho,\pi)+\ln(2/\delta)\big)}{K}+\frac{\sum_{j=1}^{J}w_j\ln\sqrt{|C_j|}}{K},
\end{aligned}$$

where the second inequality holds due to the 'change of measure' lemma (cf. Lemma 4), the third inequality uses Markov's inequality, and the last inequality proceeds as the same as the proof of Proposition 1. Further denote $\alpha_j=\frac{w_j}{\mathbf{w}}$, then we have $\sum_j\alpha_j=1$, and hence,

$$\begin{aligned}
&\frac{\sum_{j=1}^{J}w_j\ln\sqrt{|C_j|}}{K}=\frac{\mathbf{w}}{K}\sum_{j=1}^{J}\alpha_j\ln\sqrt{|C_j|}\\
\le&\frac{\mathbf{w}}{K}\ln\sum_{j=1}^{J}\frac{w_j\sqrt{|C_j|}}{\mathbf{w}} \quad \text{(Jensen)}\\
\le&\frac{\mathbf{w}}{K}\ln\frac{(\sum_{j=1}^{J}w_j)^{1/2}(\sum_{j=1}^{J}w_j|C_j|)^{1/2}}{\mathbf{w}} \quad \text{(Schwartz)}\\
=&\frac{\mathbf{w}}{K}\ln\sqrt{\frac{K}{\mathbf{w}}}.
\end{aligned}$$

Combining the above results we finally obtain

$$\begin{aligned}
&\mathrm{kl}(\mathbf{E}_\rho r(f,\mathcal{S}),\mathbf{E}_\rho R(f,\mathcal{S}))\\
\le&\frac{\mathbf{w}\big(\mathcal{K}(\rho,\pi)+\ln(2/\delta)\big)}{K}+\frac{\mathbf{w}}{K}\ln\sqrt{\frac{K}{\mathbf{w}}}.
\end{aligned}\tag{5}$$

Actually, the RHS of inequality (5) is an increasing function with respect to $\mathbf{w}$ (the detailed proof is left to readers), and hence achieve its minimum when $\mathbf{w}=\chi^*(\mathcal{S})$. Thus we complete our whole proof. ■

With almost the same proceedings as that in the proof of Propositions 2-3, we can immediately yield the following two propositions that apply to meta-learning setting where dependence exists among different samples . The detailed proof is left to readers.

Proposition 8 *For any $\delta\in(0,1)$, with probability at least $1-\delta$ over the draw of n distributions $\mathbf{D}=\{D_i\}_{i=1}^n$, the following holds for any hyper-posterior $\mathcal{Q}$:*

$$|er(\mathcal{Q})-\widetilde{er}(\mathcal{Q})|\le\sqrt{\frac{\chi^*(\mathbf{D})[\mathcal{K}(\mathcal{Q},\mathcal{P})+\ln(\frac{2}{\delta}\sqrt{\frac{n}{\chi^*(\mathbf{D})}})]}{2n}},$$

where $\chi^*(\mathbf{D})$ denotes the fractional chromatic number of the dependence graph of $\mathbf{D}$.

Proposition 9 *For any hyper-prior $\mathcal{P}$, any $\delta\in(0,1)$, with probability at least $1-\delta$ over the draw of the training sample $\mathbf{S}=\{S_i\}_{i=1}^n$, the following holds for any hyper-posterior $\mathcal{Q}$:*

$$|\widetilde{er}(\mathcal{Q})-\widehat{er}(\mathcal{Q})|\le\sqrt{\frac{2\Delta\widehat{er}(\mathcal{Q})}{mn}}+\frac{2\Delta}{mn},$$

where $\Delta=\chi^*(\mathbf{S})\big[\mathcal{K}(\mathcal{Q},\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\sum_{i=1}^n\mathcal{K}(Q_i,P)+\ln\frac{2\sqrt{mn}}{\delta\sqrt{\chi^*(\mathbf{S})}}\big]$, $\chi^*(\mathbf{S})$ denotes the fractional chromatic number of the dependence graph of $\mathbf{S}=\{S_i\}_{i=1}^n=\{z_{ij}\}_{i=1,j=1}^{n,m}$.

Proof of Theorem 6 in the main paper.
Combining Propositions 8-9 and utilizing union bound give the high-probability bound on the transfer risk $er(\mathcal{Q})$. ■

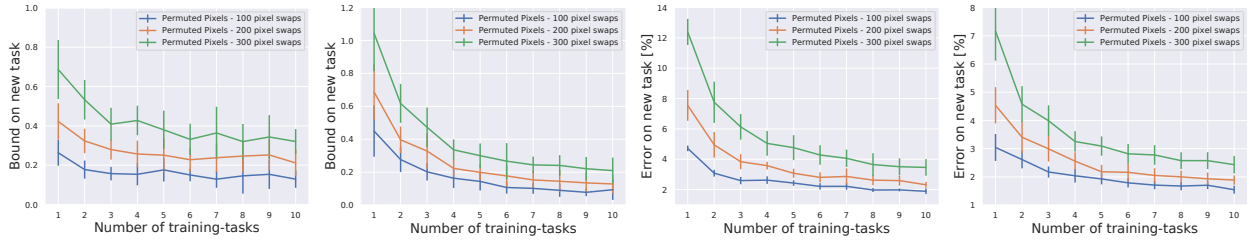
We further provide an example as follow to illustrate how to estimate the chromatic number $\chi^*(\mathbf{S})$ of the fractional cover in the dependent meta learning setting.

Example 1 *Consider a meta sample $\mathbf{S}=\{z_{ij}\}_{i,j=1}^{n,m}$, where $\forall j\in[m], z_{ij}\overset{i.i.d.}{\sim}D_i$, but there exists dependency between samples drawn from different distributions. Then we can set the exact fractional cover of $\Gamma(\mathbf{S})$ as $\{(C_j,1)\}_{j=1}^n$, and hence the chromatic number $\chi^*(\mathbf{S})\le\sum_{j=1}^n 1=n$.*

E. Details of Classification Experiments

E.1. Distributions of Neural Network

For classification problems, we can develop two meta-learning algorithms by directly setting our kl-bound and Catoni-bound as minimization objective functions. It suffices to specify: (1) the explicit form of KL-divergences in our meta-learning bounds, (2) how to approximate the expectation $P\sim\mathcal{Q}$. To tackle the first issue, we need to define



(a) kl-bound (b) Catoni-bound (c) kl-bound error (d) Catoni-bound error

Figure 2. Average test bounds and test errors of our two bound-minimizing meta-learning algorithms on new classification tasks for different numbers of training-tasks and for different pixel-shuffled environments (average over 20 meta-test tasks). (a)-(b): Test bounds of kl-bound and Catoni-bound. (c)-(d): Test errors of kl-bound and Catoni-bound.

the distribution of the hypothesis class $\mathcal{H} = \{w : w \in \mathbb{R}^d\}$ (i.e., neural networks), where d is the dimension of the parameter w. As that in previous work (Amit & Meir, 2018), we first set both the hyper-prior and hyper-posterior over $\mathcal{M}_1(\mathcal{H})$ as isotropic Gaussian:

$$\mathcal{P} = \mathcal{N}(0, \kappa_{\mathcal{P}}^2 I_{d\times d}), \mathcal{Q}_\theta = \mathcal{N}(\theta, \kappa_{\mathcal{Q}}^2 I_{d\times d}),$$

where $\kappa_{\mathcal{P}}, \kappa_{\mathcal{Q}} > 0$ are both predefined constants, θ is the optimization parameter. Then the KL-divergence between $\mathcal{Q}_\theta$ and $\mathcal{P}$ can be calculated as

$$\mathcal{K}(\mathcal{Q}_\theta, \mathcal{P}) = \frac{||\theta||_2^2 + \kappa_{\mathcal{Q}}^2}{2\kappa_{\mathcal{P}}^2} + \ln\frac{\kappa_{\mathcal{P}}}{\kappa_{\mathcal{Q}}} - \frac{1}{2}. \tag{6}$$

Next, we consider the form of prior and posterior over the hypothesis space $\mathcal{H}$. We define the prior P_θ and the posteriors Q_{ϕ_i} ($\phi_i \in \mathbb{R}^d$ is the hyperparameter) as the factorized Gaussian distributions for computational convenience:

$$P_\theta(w) = \prod_{k=1}^{d} \mathcal{N}(w_k; \mu_{P,k}, \sigma_{P,k}^2),$$
$$Q_{\phi_i}(w) = \prod_{k=1}^{d} \mathcal{N}(w_k; u_{i,k}, \sigma_{i,k}^2),$$

where $\theta = (\mu_P, \rho_P) \in \mathbb{R}^{2d}$ is composed of the means $\mu_{P,k}$ and log-variances of each weight $\rho_{P,k} = \ln\sigma_{P,k}^2, k \in [d]$. The posterior vectors $(\mu_i, \rho_i) \in \mathbb{R}^{2d}$ has a similar structure. Then the KL-divergence $\mathcal{K}(Q_{\phi_i}, P_\theta)$ can be calculated as:

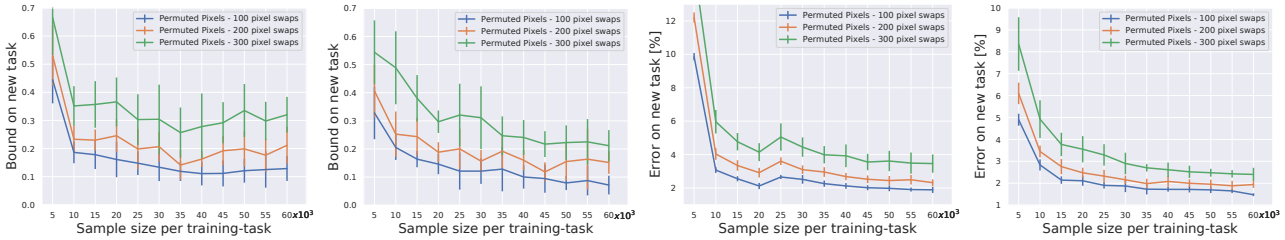
$$\frac{1}{2}\sum_{k=1}^{d}\Big\{\ln\frac{\sigma_{P,k}^2}{\sigma_{i,k}^2} + \frac{(\sigma_{i,k}^2 + (\mu_{i,k} - \mu_{P,k})^2)}{\sigma_{P,k}^2} - 1\Big\}. \tag{7}$$

Secondly, to approximate the expectation $P \sim \mathcal{Q}$ in our kl-bound and Catoni-bound, we utilize the Monte-Carlo method. Concretely, we calculate the expectations by averaging several Monte-Carlo samples of P and adding Gaussian noise to the parameter θ during the meta-training process: $\tilde{\theta} = \theta + \epsilon$, $\epsilon \sim \mathcal{N}(0, \kappa_{\mathcal{Q}}^2 I_{d\times d})$. Therefore from the above discussions, we can derive two PAC-Bayesian algorithms for meta-learning classification problems with deep neural networks. The detailed pseudo code of setting our PAC-Bayesian Catoni-bound as training objective is shown in Algorithm 1, where we set $C_1 = 2, C_2 = 3$ for convenience. The pseudo code of minimizing our kl-bound can be illustrated in a similar way. In practice, we set the parameters $\kappa_{\mathcal{P}} = 2000$ and $\kappa_{\mathcal{Q}} = 0.001$ respectively, and the confidence parameter $\delta = 0.1$. ADAM is chose as the optimizer with learning rate of 10^{-3} for all experiments.

During the meta-test phase, the informative prior is sampled randomly from the learned hyper-posterior $\mathcal{Q}_\theta$, and we take this prior and the scarce data of the novel task as input to learn a posterior. The test bound is calculated on the

Algorithm 1 Catoni-bound-minimizing meta-learning algorithm (meta-training phase)

Input: Datasets from n training tasks: $S_1, ..., S_n$.
Output: Parameters θ of hyper-posterior $\mathcal{Q}_\theta$.
Initialize:
$\theta = (\mu_P, \rho_P) \in \mathbb{R}^{2d}$, $\phi_i = (\mu_i, \rho_i) \in \mathbb{R}^{2d}, i = 1, ..., n$.
while *not converged* **do**
  for $i \in \{1, ..n\}$ **do**
    Sample a mini-batch S_i' from datasets S_i.
    Calculate $\mathbf{E}_{P_\theta\sim\mathcal{Q}_\theta}\widehat{er}(Q_i, S_i)$ with the mini-batch S_i' by averaging Monte-Carlo draws.
    Calculate $\mathcal{K}(\mathcal{Q}_\theta, \mathcal{P})$ with Eq. (6).
    Calculate $\mathbf{E}_{P_\theta\sim\mathcal{Q}_\theta}\mathcal{K}(Q_{\phi_i}, P_\theta)$ with Eq. (7) by averaging Monte-Carlo draws.
  end for
  Calculate the meta-training Catoni-bound in Eq. (4) with $\mathbf{E}_{P_\theta\sim\mathcal{Q}_\theta}\widehat{er}(Q_i, S_i)$, $\mathcal{K}(\mathcal{Q}_\theta, \mathcal{P})$ and $\mathbf{E}_{P_\theta\sim\mathcal{Q}_\theta}\mathcal{K}(Q_{\phi_i}, P_\theta), i = 1, ..., n$.
  Calculate the gradient of Catoni-bound w.r.t $\{\theta, \phi_1, ..., \phi_n\}$ using backpropagation.
  Take an optimization step.
end while
Return θ



(a) kl-bound (b) Catoni-bound (c) kl-bound error (d) Catoni-bound error

Figure 3. Average test bounds and test errors of our two bound-minimizing meta-learning algorithms on new classification tasks for different sample sizes per training-task and for different pixel-shuffled environments (average over 20 meta-test tasks). (a)-(b): Test bounds of kl-bound and Catoni-bound. (c)-(d): Test errors of kl-bound and Catoni-bound.

novel task by using the PAC-Bayesian bound for single-task learning as the minimization objective (i.e., only calculate the task-level complexity in the meta-training bound).

E.2. Convergence Analysis of PAC-Bayesian kl-bound and Catoni-bound for Meta-Learning

In this subsection, we provide visualization of the convergence performance of our kl-bound and Catoni-bound. Such experiment is conducted for a large range of the number n of training tasks and a large range of the sample size m per task in different classification environments. Detailed information can be found in Figure 2 & Figure 3. We can observe that: (1) With the increase of the number of meta-training tasks, our meta-learning algorithms by minimizing the proposed kl-bound and Catoni-bound can achieve lower test bounds and lower test errors over the novel task. This verifies the asymptotic behaviour of our two meta-learning bounds. (2) When the number of training task or the sample size per task is rather small (i.e., $n = 1$ or $m = 5,000$), both our kl-bound and Catoni-bound suffer performance degradation. However, when $n \geq 2$ or $m \geq 10,000$, our two bound-minimizing algorithms obtain much better performance. This indicates the value of extracting knowledge from other similar training tasks that have sufficient training data. (3) Our Catoni-bound can always achieve a lower level than kl-bound, in terms of the test bound and the test error, which is consistent with the tightness of the Catoni-bound.

F. Details of Regression Experiments

Table 4. The number n of meta-training tasks and the sample size m per task in different regression environments.

	Cauchy	SwissFEL	Physionet (GCS-HCT)	Berkeley
n	20	5	100	36
m	20	200	4 - 24	288

F.1. Regression Environments

We provide more details on the five regression environments in Table 4. The information includes the number of training tasks and the sample size per task of different environments. The comprehensive introduction of each environment can be found in (Rothfuss et al., 2021) [Appendix E].

F.2. Approximating Gibbs Optimal Hyper-posterior with SVGD Inference Method

Algorithm 2 GOHP with SVGD approximation of $\mathcal{Q}^*$ (meta-training phase)

Input: Hyper-prior $\mathcal{P}$, datasets $S_1, ..., S_n$.
Hyper-parameter: SVGD kernel function $k(\cdot,\cdot)$, step size η, scaler factor β in Eq. (8).
Output: Set of priors $\{P_{\phi_1}, ..., P_{\phi_K}\}$.
Initialize: $\boldsymbol{\phi} := [\phi_1, ..., \phi_K]$, with $\phi_k \sim \mathcal{P}$.
while not converged **do**
  for $k = 1, ..., K$ **do**
    for $i = 1, ..., n$ **do**
      $\ln Z_{i,k} \leftarrow \texttt{MLL\_Estimator}(S_i, P_{\phi_k})$
    end for
    $\nabla_{\phi_k} \ln \tilde{\mathcal{Q}}^* \leftarrow \nabla_{\phi_k} \ln \mathcal{P} + \frac{\beta}{nm} \sum_{i=1}^{n} \nabla_{\phi_k} \ln Z_{i,k}$
  end for
  $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \eta\, \mathbf{K}\, \nabla_{\boldsymbol{\phi}} \ln \tilde{\mathcal{Q}}^* + \nabla_{\boldsymbol{\phi}} \mathbf{K}$ // SVGD update
end while
Return $\{P_{\phi_1}, ..., P_{\phi_K}\}$

In this subsection, we detail how to employ the inference method SVGD (Liu & Wang, 2016) to approximate the Gibbs optimal hyper-posterior $\mathcal{Q}^*$. We borrow the idea from (Rothfuss et al., 2021) to develop our GOHP algorithm. Concretely, SVGD approximates $\mathcal{Q}^*$ as a set of particles $\hat{\mathcal{Q}} = \{P_{\phi_1}, \ldots, P_{\phi_K}\}$, where P_ϕ represents a prior with parameter ϕ. Initially, we sample K particles ϕ_k from the hyper-prior $\mathcal{P}$. Then according to the explicit form of $\mathcal{Q}^*$ in Corollary 2 of the main paper, we can compute the gradient of $\mathcal{Q}^*$ w.r.t. the parameters $\phi_k, k \in [K]$:

$$\nabla_{\phi_k} \ln \mathcal{Q}^*(\phi_k) = \nabla_{\phi_k} \ln \mathcal{P}(\phi_k) + \frac{\beta}{nm} \sum_{i=1}^{n} \nabla_{\phi_k} \ln Z(S_i, P_{\phi_k}), \tag{8}$$

where the marginal log-likelihood (MLL) $\ln Z(S_i, P_{\phi_k})$ is

Table 5. Different PAC-Bayesian meta-learning bounds on $er(\mathcal{Q})$. **Bound = Empirical Error + Environment-level Complexity + Task-level Complexity**. n is the number of training tasks, m is the size of each training dataset S_i ($i \in [n]$). $\mathcal{P}, \mathcal{Q} \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{H}))$ are hyper-prior and hyper-posterior respectively. $P, Q_i = Q(S_i, P) \in \mathcal{M}_1(\mathcal{H})$ are the prior and the posterior for the i-th training task. Our Catoni-bound holds for any positive constant $C_1, C_2 > 1$.

Classical Bounds	Empirical Error	Environment-level Complexity	Task-level Complexity
Pentina & Lampert	$\widehat{er}(\mathcal{Q})$	$\frac{1}{\sqrt{n}}\big(\mathcal{K}(\mathcal{Q},\mathcal{P}) + \frac{1}{8} + \ln\frac{\delta}{2}\big)$	$\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})+\sum_{i=1}^n \mathbf{E}_{P\sim\mathcal{Q}}\mathcal{K}(Q_i,P)}{n\sqrt{m}} + \frac{1}{8\sqrt{m}} + \frac{1}{n\sqrt{m}}\ln\frac{2}{\delta}$
Amit & Meir	$\widehat{er}(\mathcal{Q})$	$\sqrt{\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})+\ln\frac{2n}{\delta}}{2(n-1)}}$	$\frac{1}{n}\sum_{i=1}^n \sqrt{\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\mathcal{K}(Q_i,P)+\ln(2nm/\delta)}{2(m-1)}}$
Rothfuss et al.	$\widehat{er}(\mathcal{Q})$	$\frac{1}{\sqrt{n}}\mathcal{K}(\mathcal{Q},\mathcal{P}) + \frac{1}{8\sqrt{n}}$	$\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})+\sum_{i=1}^n \mathbf{E}_{P\sim\mathcal{Q}}\mathcal{K}(Q_i,P)}{n\sqrt{m}} + \big(\frac{1}{8n\sqrt{m}} + \frac{1}{\sqrt{n}}\ln\frac{1}{\delta}\big)$
kl-bound (ours)	$\widehat{er}(\mathcal{Q})$	$\sqrt{\frac{\mathcal{K}(\mathcal{Q},\mathcal{P})+\ln\frac{2\sqrt{n}}{\delta}}{2n}}$	$\sqrt{\frac{2\Delta\widehat{er}(\mathcal{Q})}{mn}} + \frac{2\Delta}{mn}, \Delta = \mathcal{K}(\mathcal{Q},\mathcal{P}) + \mathbf{E}_{P\sim\mathcal{Q}}\sum_{i=1}^n \mathcal{K}(Q_i,P) + \ln\frac{2\sqrt{mn}}{\delta}$
Catoni-bound (ours)	$\frac{C_1C_2\ln C_1\ln C_2}{(C_1-1)(C_2-1)}\widehat{er}(\mathcal{Q})$	$\frac{C_1\big(\mathcal{K}(\mathcal{Q},\mathcal{P})+\ln(2/\delta)\big)}{n(C_1-1)}$	$\frac{C_1C_2\ln C_1\big(\mathcal{K}(\mathcal{Q},\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\sum_{i=1}^n \mathcal{K}(Q_i,P)+\ln(2/\delta)\big)}{(C_1-1)(C_2-1)nm}$

approximated by numerical Monte Carlo estimates. Then we update the particles with the SVGD update rule:

$$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \eta\, \mathbf{K}\, \nabla_{\boldsymbol{\phi}} \ln\tilde{\mathcal{Q}}^* + \nabla_{\boldsymbol{\phi}}\mathbf{K},$$

where $\boldsymbol{\phi} = [\phi_1, ..., \phi_K]^\top$ is the stacked particles matrix, $\nabla_{\boldsymbol{\phi}}\ln\tilde{\mathcal{Q}}^* = [\nabla_{\phi_1}\ln\mathcal{Q}^*(\phi_1), ..., \nabla_{\phi_K}\ln\mathcal{Q}^*(\phi_K)]^\top$ the stacked matrix of gradients, $\mathbf{K} = [k(\phi_k, \phi_{k'})]_{k,k'}$ the kernel matrix induced by the kernel function $k(\cdot,\cdot)$ and η the step size for updates. The Pseudo code for meta-training can be found in Algorithm 2.

Algorithm 3 GOHP with SVGD on the novel tasks (meta-test phase)

Input: Set of priors $\{P_{\phi_1}, ..., P_{\phi_K}\}$, dataset S_{n+1} from novel task.
Hyper-parameter: Kernel function $k(\cdot,\cdot)$, SVGD step size η, number of particles L.
Output: A set of neural networks parameters $\bigcup_{k=1}^K\{\theta_1^k ..., \theta_L^k\}$.
for $k = 1, ..., K$ **do**
  Initialize $\{\theta_1^k, ..., \theta_L^k\}$, $\theta_l^k \sim P_{\phi_k}$, $l \in [L]$.
  while not converged **do**
    for $l = 1, ..., L$ **do**
      $\nabla_{\theta_l^k}\ln Q^*(\theta_l^k) \leftarrow \nabla_{\theta_l^k}\ln P_{\phi_k} - m\nabla_{\theta_l^k}\widehat{er}(h, S_{n+1})$
    end for
    $\theta_l^k \leftarrow \theta_l^k + \frac{\eta}{L}\sum_{l'=1}^L [k(\theta_{l'}^k, \theta_l^k)\nabla_{\theta_{l'}^k}\ln Q^*(\theta_{l'}^k) + \nabla_{\theta_{l'}^k} k(\theta_{l'}^k, \theta_l^k)]$.
    // SVGD update
  end while
end for
Return $\bigcup_{k=1}^K\{\theta_1^k ..., \theta_L^k\}$

F.3. Applying Gibbs Optimal Hyper-posterior to Novel Tasks

During the meta-test phase, the extracted knowledge from the n training tasks is now applied to a novel task by a base learner. The base learner is given a training dataset $S_{n+1} \sim D_{n+1}$, where D_{n+1} is sampled from the same environment τ. We still use statistical inference method to approximate the Gibb optimal posterior $Q^*(S_{n+1}, P)$ defined in Corollary 1 of the main paper. Then we use the SVGD update rule in Eq. (8) to update the parameter θ of the neural networks for the novel task. Algorithm 3 gives the steps of the training procedure during the meta-test phase.

For a data point x^* from the held-out evaluation set $S^* \sim D_{n+1}$, the neural network predictor outputs a probability distribution as $\hat{p}(y^*|x^*, S_{n+1}) \leftarrow \frac{1}{KL}\sum_{k,l=1}^{K,L} p(y^*|h_{\theta_l^k}(x^*))$. Then the mean prediction is set as the expectation of the distribution $\hat{p}$, i.e., $\hat{y} = \mathbf{E}_{y^*\sim\hat{p}}(y^*|x^*, S_{n+1})$. Thus the root mean squared error (RMSE) over the novel task D_{n+1} is calculated as follow:

$$RMSE = \sqrt{\frac{1}{|S^*|}\sum_{(x^*,y^*)\in S^*}(y^* - \hat{y})^2}.$$

Remark 2 *Note that our mete regression algorithm GOHP achieves analogous experimental results w.r.t. the latest PACOH algorithm on the regression datasets in Table 3 in the main paper. There are two reasons for GOHP's analogous performance w.r.t. PACOH:* ***(1)*** *Both algorithms minimize similar objectives: although our bound of $O\big((\frac{1}{n} + \frac{1}{mn})\mathcal{K}(\mathcal{Q},\mathcal{P})\big)$ in Corollary 1 is sharper than that of $O\big((\frac{1}{n} + \frac{1}{mn})\mathcal{K}(\mathcal{Q},\mathcal{P}) + \theta\big)$ in (Rothfuss et al., 2021)[Corollary 1], where θ is a constant that will not decrease with the increase of the number n of training tasks or the sample size m per task, Rothfuss's PACOH omits the constant term θ during the optimization process. Thus, both our minimization objective and that of PACOH are almost the same, except for the multiplicative factor before $\mathcal{K}(\mathcal{Q},\mathcal{P})$.* ***(2)*** *Both algorithms use the same approximation technique: both GOHP and PACOH employ the same inference method SVGD (Liu & Wang, 2016) to minimize $O\big((\frac{1}{n} + \frac{1}{mn})\mathcal{K}(\mathcal{Q},\mathcal{P})\big)$ and update the hyper-posterior $\mathcal{Q}$ iteratively.*

Table 6. Notations of PAC-Bayesian single-task learning and PAC-Bayesian meta-learning. $\mathcal{H}$ is the hypothesis space, and $l : \mathcal{H} \times \mathcal{Z} \to [0, 1]$ is the bounded loss function. In the main paper, we write $\widehat{er}(Q) \triangleq \widehat{er}(Q, S)$, $er(\mathcal{Q}) \triangleq er(\mathcal{Q}, \tau)$ for abbreviation.

PAC-Bayesian Single-Task Learning		PAC-Bayesian Meta-Learning	
Sample	$z \in \mathcal{Z}$	Sample	$S \in \mathcal{Z}^m$
Training Set	$S = \{z_i\}_{i=1}^m \in \mathcal{Z}^m$	Training Set	$\mathbf{S} = \{S_i\}_{i=1}^n \in (\mathcal{Z}^m)^n$
Unknown Task	$D \in \mathcal{M}_1(\mathcal{Z})$	Unknown Environment	$\tau \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{Z}))$
Prior	$P \in \mathcal{M}_1(\mathcal{H})$	Hyper-Prior	$\mathcal{P} \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{H}))$
Posterior	$Q \in \mathcal{M}_1(\mathcal{H})$	Hyper-Posterior	$\mathcal{Q} \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{H}))$
Empirical Error	$\widehat{er}(Q, S) = \mathbf{E}_{h\sim Q} \frac{1}{m} \sum_{i=1}^m l(h, z_i)$	Empirical Error	$\widehat{er}(\mathcal{Q}, \mathbf{S}) = \mathbf{E}_{P\sim\mathcal{Q}} \frac{1}{n} \sum_{i=1}^n \widehat{er}(Q(S_i, P), S_i)$
Expected Error	$er(Q, D) = \mathbf{E}_{h\sim Q}\mathbf{E}_{z\sim D} l(h, z)$	Transfer Error	$er(\mathcal{Q}, \tau) = \mathbf{E}_{P\sim\mathcal{Q}}\mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} er(Q(S, P), D)$

G. Explicit Forms of Different PAC-Bayesian Meta-Learning Bounds

In this section, we provide the explicit forms of different PAC-Bayesian bounds for meta-learning in Table 5. Note that Table 5 provides the detailed version of those PAC-Bayesian bounds listed in Table 1 of the main paper.

Remark 3 *We remark that it is infeasible to directly compare our results with the latest generalization bounds in (Farid & Majumdar, 2021) which combines PAC-Bayes analysis and algorithmic stability theory. The reasons lie in two aspects:* ***(1)*** *It is hard to compute a precise PAC-Bayes bound in (Farid & Majumdar, 2021)[Theorem 3], as the uniform stability parameter (referred as β) in their bound depends on L-Lipschitzness and B-smoothness of neural networks, where the constants L and B are always unknown and truly big ($>> 1$).* ***(2)*** *Nevertheless, we can make a rough comparison as follow. In the convex case, the stochastic gradient descent (SGD) algorithm with constant step sizes has a stability parameter $\beta = O(L^2T/m)$, where T is the total number of iteration. Recall that $T = 200$, $m = 2,000$ in our meta-test phase, so $\beta >> 200/2000 = 0.1$ and is larger than our reported Catoni-bound in Table 2. Therefore, Farid's generalization bound in the convex case is not tight enough, let alone the bound in the non-convex case (e.g. deep neural networks).*

H. Notations of Single-Task Learning and Meta-Learning

In this section, we provide notations of PAC-Bayesian single-task learning and PAC-Bayesian meta-learning in Table 6 for readers convenient reference.